Universität Bielefeld

CITEC

# RDM + Conquaire

## RDM: A library perspective of versioning, curating and archiving research data from diverse domains

**VID AYER**
**Scientific Researcher, CITEC,**
**Bielefeld University, Germany**

Talk @ DI4R
09-Oct-2018, Lisbon, Portugal.

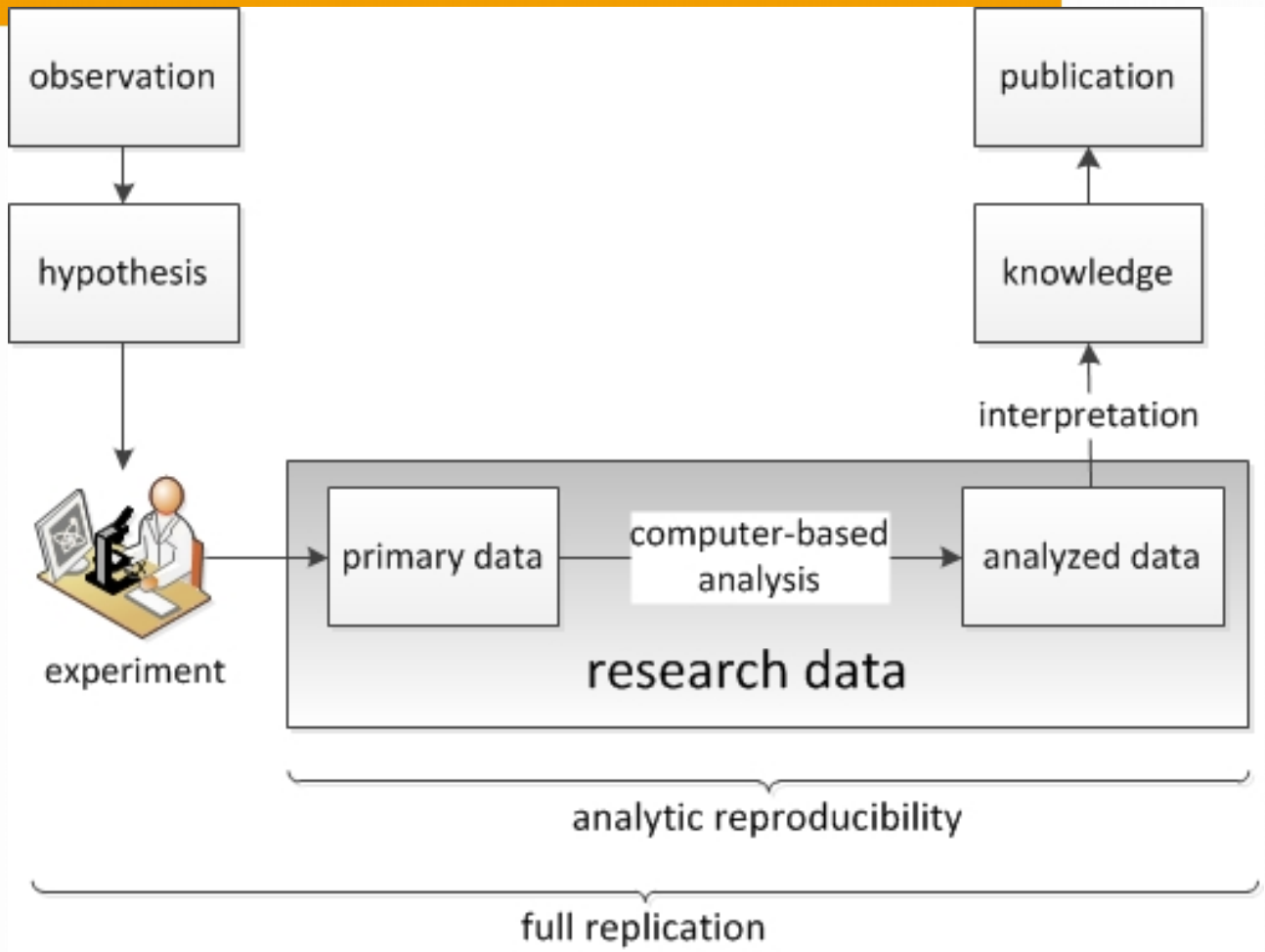
# Agenda

- Conquaire Introduction
- Conquaire & computational reproducibility
- Library Infrastructure - RDM
- RDM => Conquaire (Gitlab + CI) & PUB

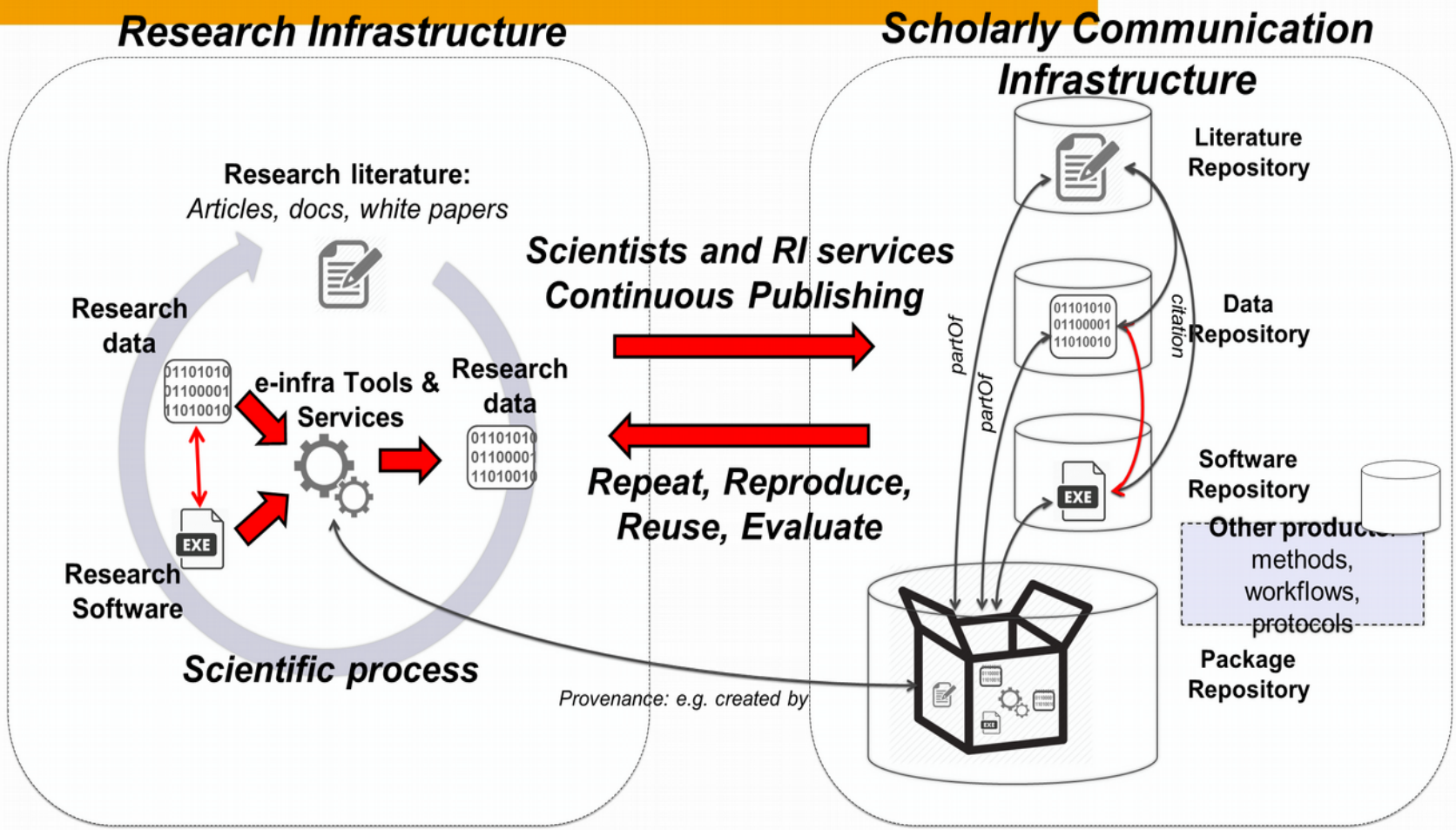
# About

- **DFG funded:** 2016 – 2019.
- **CITEC + Bielefeld University Library**
- **9 research groups:** Interdisciplinary + InterUniversity
- **Disciplines**: Applied Computational Linguistics, Biology, Computer Science, Chemistry, Economics, Linguistics, Neurobiology, Psychology, Sports Science
- **Research Data:** High Diversity (data formats, experiment tools, software)
- **DMP**: Data Management Plan

# Computational Reproducibility



# RDM



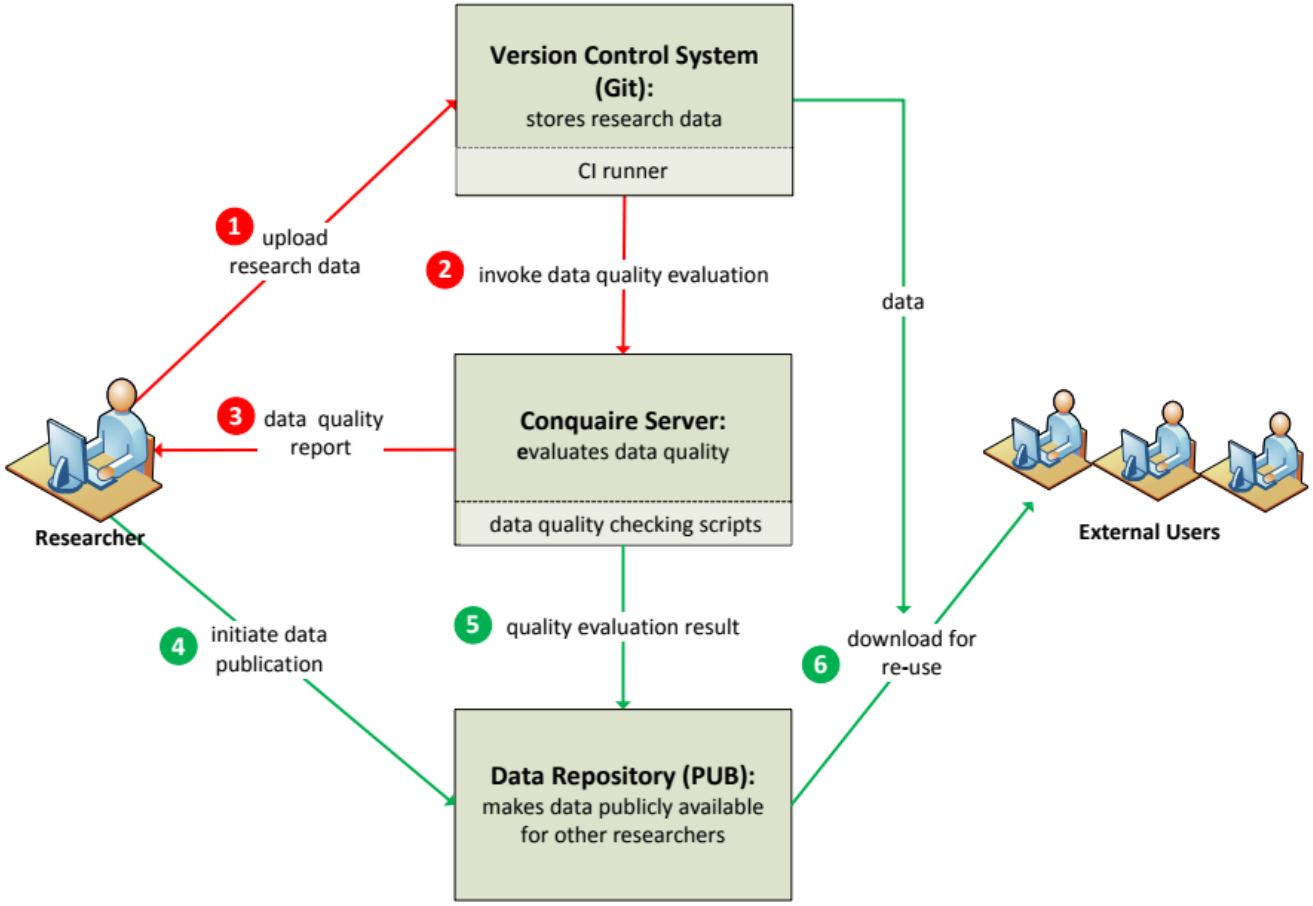
# RDM Goals

- **Research Data Management System** (RDMS): generic infrastructure, data publication in PUB
- **RDM of diverse resources:**
  - papers, manuscripts, articles
  - Research datasets = data + images+ software
  - Backend: Research Data versioned in Gitlab
- **Research Data Quality** --> ConQuaire

# RDM : Infrastructure Components

- Research Objects : Technical + Social
  - Technical aggregation of resources
  - REST(ful) API: Inclusion of publication lists
  - Record best practices and support reproducibility
  - Ontologies (Metadata): annotations
  - SRU + MODS: create your own frontends – search & retrieval via URL
- Data pipeline – FAIR principles
  - Data preservation - Citable artifacts
  - Automated checks for data (BigData)
  - Interoperability checks

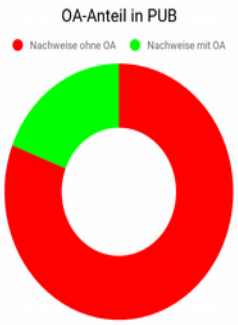
# Conquaire Architecture



# PUB !

- **Management of Institutional research output:**
  - Scientific literature + Research Data linking at #UniBi
- **Built with LibreCat:**
  - Joint effort of Lund, Gent, Bielefeld libraries.
- **Supports:**
  - Author publication lists
  - Mints DOI / URN for permanent, reliable citation
  - Interfaces (OAI, SRU, CQL)
  - Formats (DC, MODS, DataCite, XmetaDissPlus)
- **59,564 publication references: ~19% OA**
- **3,919 pers. Publication lists**
- **1.9 million views (2017)**
- **> 900,000 downloads (2017)**
- **> 12,500 publication references with an ORCID-iD: (> 430 scientists with an ORCID-iD)**

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
http://doi.org/10.17616/R3W60X
PUB Data Publications



# DIRA: Data IRreproducibility Analyzer

- **Generic quality checks**
- **Implemented CSV file testing:**
  - Eg. declare dtype in format file to process data types.
- **Data Quality checks - computational reproducibility**
- **Ensure data reusability**
- **Continuous Integration (CI) support**

# Data Diversity Challenges

- **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..
- **CI Maintenance:**
  - Costs to maintain infrastructure
  - FOSS (Free & Open Source Software) easier to maintain
  - ‘Non-open’ software costs more – versioning, licence restrictions

# Computational Reproducibility Challenges!

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

# Gitlab-CI

- **CI standardizes technology**
  - Platform
  - Tools
  - Enhances cross-domain data interoperability - RDM service
- **Automated Quality Checking Tool**
  - .CSV file checking - tested & implemented
  - .XML file checking - WIP

# Gitlab.UB

- **Collaboration tool:**
  - Scientists & researchers across projects
  - Teaching tool – lecturers
  - Students use GitLab
- **Most active user: Digital humanities project**
  - Luhmann co-operative effort + Cologne University
  - Annotate digitized index cards - Niklas Luhmann
  - Based on XML language TEI
- **412 active users in 68 groups - created 641 projects**

# CaseStudy: Psycholinguistics

- **Manuscript (Accepted)**: Evidence for early comprehension of action verbs
- **Toolkit**: Python-2.x, ported to 3.6, Pandas, Matplotlib
- **Curated digital dataset**: Computationally Reproducible
  - Raw data: children (9-10 month) audio/ videos (private)
  - Gaze data (semi-processed data): looking time, stored in .CSV format
  - Scripts, Data Visualisation (IPython notebooks) scripts, Docs
  - Generic CI pipeline: Data Visualisation & .CSV files
  - PUB: DOI, links to download
- **Users**:
  - HTML & text logs
  - Notifications – data changes
  - DOI for publications

# Gitlab + PUB : Example

A C++ Implementation of the reversed Attentional Vector Sum (rAVS) model

Download ZIP-Archive | TAR-Archive
DOI 10.5072/test/2737400
*Software*

Details Versions

## Versions of this Dataset

- Version from 2018-02-14 12:45:10
  DOI: 10.5072/test/2737400/cb6381c8
  View files on GitLab@UniBi
  ZIP-Archive | TAR-Archive
- Version from 2018-01-26 11:52:57
  DOI: 10.5072/test/2737400/d0be1dc3
  View files on GitLab@UniBi
  ZIP-Archive | TAR-Archive
- Version from 2017-05-29 19:18:42
  DOI: 10.5072/test/2737400/b7441f81
  View files on GitLab@UniBi
  ZIP-Archive | TAR-Archive

# PUB : Example

## A C++ Implementation of the reversed Attentional Vector Sum (rAVS) model

| | |
|---|---|
| Download | ZIP-Archive \| TAR-Archive |
| DOI | 10.5072/test/2737400 |
| *Software* | |

**Details** **Versions**

| | |
|---|---|
| Creator | Kluth, T |
| Department | **Department of Physics -> Relativity Theory Group** |
| Publishing Year | **2016** |
| PUB-ID | **2737400** |

## Cite this

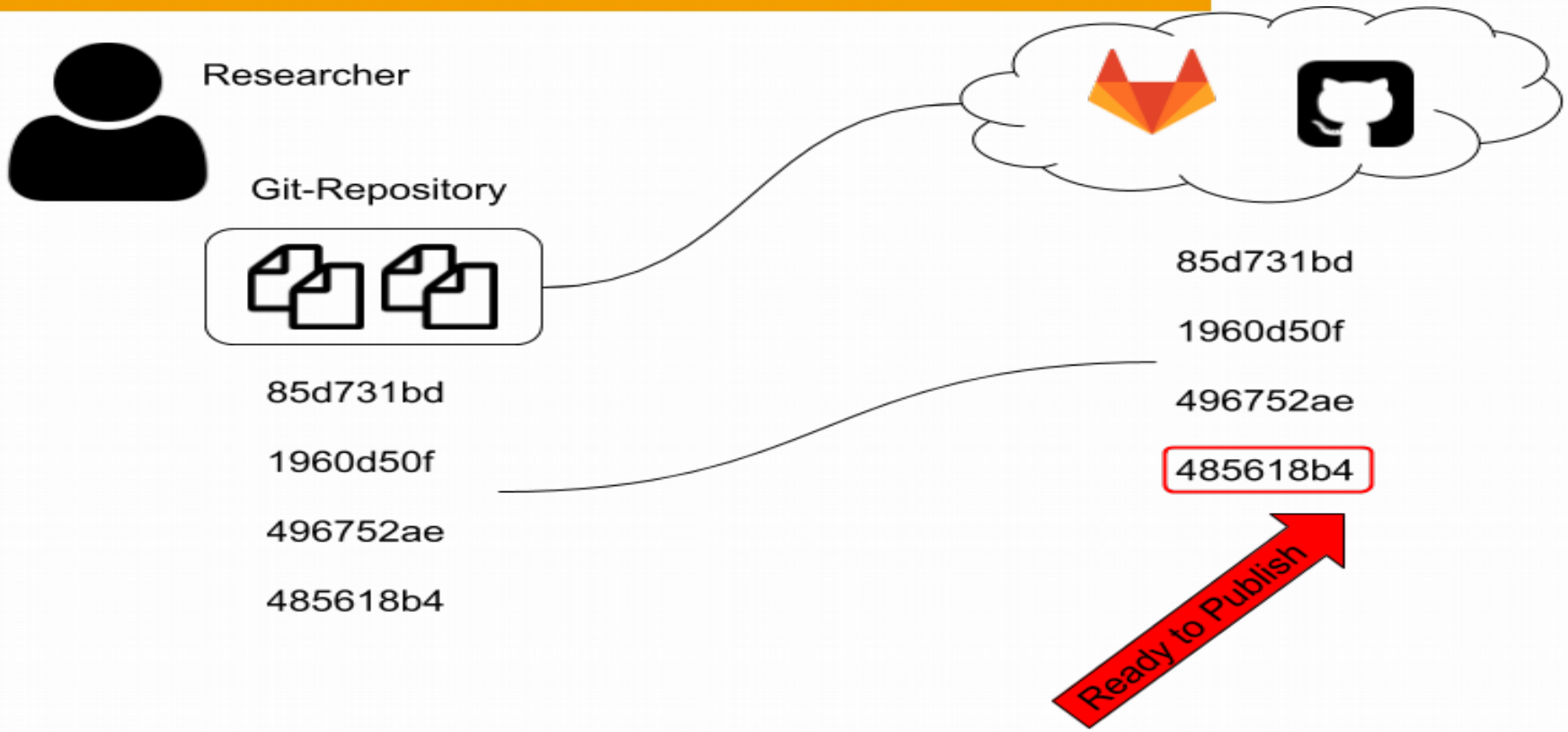
# PUB : Dataset Version

Details | Versions

## Versions of this Dataset

- Version from 2018-02-14 12:45:10
  DOI: **10.5072/test/2737400/cb6381c8**
  **View files on GitLab@UniBi**
  **ZIP-Archive | TAR-Archive**

# Gitlab Versioning




Icons made by Dave Gandy (https://www.flaticon.com/authors/dave-gandy), CC 3.0 BY

# PUB : Dataset Version

PUBLIKATIONEN UNIVERSITÄT BIELEFELD

85d731bd

1960d50f

496752ae

485618b4

ID: 1234
Title: ....
Author: ...
GitLab-URL:
https://gitlab..../485618b4
DOI:
10.10572/pub/1234
etc.

| | |
|---|---|
| 10.10572/pub/1234/485618b4 | 10.10572/pub/1234 |
| Version-DOI | Collection-DOI |

DataCite
FIND, ACCESS, AND REUSE DATA

Icons made by Dave Gandy (https://www.flaticon.com/authors/dave-gandy), CC 3.0 BY

# Thank You!

- **Questions?**
- **Contact:**
  - Email: **ayer@uni-bielefeld.de**
  - Twitter: @svaksha
  - Website: **http://conquaire.uni-bielefeld.de**
  - Github: https://github.com/svaksha