Sustainable and open ocean research software

Prof. Dr. Wilhelm (Willi) Hasselbring
http://se.informatik.uni-kiel.de/
Abgrenzung zu Forschungsdaten

• Software is **executable**,  
  – data is not.
• Data provides **evidence**,  
  – software provides tools.
• Software is a **creative work**,  
  – data are facts or observations.
• Software is updated more **frequently** than papers or data.

Source: Daniel S. Katz et al.  
https://github.com/danielskatz/software-vs-data
Relevance of Research Software

What would happen if you could not longer use research software?

- Research would be impossible: 69%
- Possible, but difficult: 21%
- No effect: 10%

https://www.software.ac.uk/blog/2016-07-26-its-impossible-conduct-research-without-software-say-7-out-10-uk-researchers
Gene name errors are widespread in the scientific literature

Mark Ziemann\(^1\), Yotam Eren\(^{1,2}\) and Assam El-Osta\(^{1,3}\)*

**Abstract**

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

**Keywords:** Microsoft Excel, Gene symbol, Supplementary data

**Abbreviations:** GEO, Gene Expression Omnibus; JIF, journal impact factor
Research Software Sustainability

Without software, modern research would not be possible. This report recommends practices of software sustainability to minimise the risks of reliability and reproducibility of research.

1 October 2015 - 3 March 2016, 00:00 - 00:00, Berlin, Germany

http://www.knowledge-exchange.info/event/software-sustainability
Was tun andere?
(hier UK)

https://www.software.ac.uk/

http://wssspe.researchcomputing.org.uk/
Nachhaltigkeit von Forschungssoftware
Relevanz aus Sicht der DFG als Forschungsförderer

Nachnutzung

Nachvollziehbarkeit

Nachhaltigkeit

Infrastrukturen

Reproduzierbarkeit

Quelle: Dr. Matthias Katerbow (DFG LIS): „Nachhaltigkeit von Forschungssoftware“, Helmholtz Open Science Workshop, Dresden, 2016
Nachnutzung von Forschungssoftware

• Schaffung von Rahmenbedingungen
  – Bereitstellung, Archivierung, Versionierung, etc.
  – Nutzbarhaltung, Wartung, Support
  – Qualitätssicherung
  – Lizenzmodelle

• Herausforderungen
  – Geschäftsmodelle
  – Langfristige Finanzierung
  – Integration von Weiterentwicklungen und Anpassungen durch Dritte

Quelle: Dr. Matthias Katerbow (DFG LIS)
Nachvollziehbarkeit der Methodik

- Wissenschaftliche Ergebnisse können nur reproduziert werden, wenn alle wichtigen Schritte vollständig nachvollziehbar sind
  - Was gilt hier für die Verwendung von Forschungssoftware?
    - Verfügbarkeit der Software?
    - Verfügbarkeit und Dokumentation des Quellcodes?
    - Parametrisierung der Software während der Nutzung?
    - Anpassungen der Software?

- Nachweise eigener und fremder Vorarbeiten (Zitate)
  - Was gilt hier für die Verwendung von Forschungssoftware?
    - Wie soll Software zitiert werden? (Standards z.B. zur executable citation)
    - Wie werden Anwendung und Parametrisierung von Software dokumentiert?

Quelle: Dr. Matthias Katerbow (DFG LIS)
Reproduzierbarkeit mit und durch Forschungssoftware

Quelle: Dr. Matthias Katerbow (DFG LIS)
Infrastruktur zur Nachhaltigkeit von Forschungssoftware

Herausforderungen u.a.:

• Nachhaltige Software-Entwicklung
• Nachweissysteme, Metadaten
• Zitierbarkeit
  – Erste Ansätze z.B.: GitHub Release via Zenodo archivieren

• Reproduzierbarkeit
• Geschäftsmodelle

Quelle: Dr. Matthias Katerbow (DFG LIS)
“Replication is the ultimate standard by which scientific claims are judged.”
The case for open computer programs

Darrel C. Ince¹, Leslie Hatton² & John Graham-Cumming³

• “We argue that, with some exceptions, anything less than the release of source programs is **intolerable** for results that depend on computation.

• The vagaries of hardware, software and natural language will always ensure that exact **reproducibility remains uncertain**, – but withholding code increases the chances that efforts to reproduce results will fail.”
“Science advances faster when we can build on existing results, and when new ideas can easily be measured against the state of the art.”

*Repeatability*, not necessarily *reproducibility*

Several ACM SIGMOD, SIGPLAN, and SIGSOFT conferences have initiated *artifact evaluation* processes.
What are we doing?

Hierarchical Software Landscape Visualization for System Comprehension: A Controlled Experiment

Florian Fittkau, Alexander Krause, and Wilhelm Hasselbring
Software Engineering Group, Kiel University, Kiel, Germany
Email: {ffi, akr, wha}@informatik.uni-kiel.de

Exploring Software Cities in Virtual Reality
Florian Fittkau, Alexander Krause, and Wilhelm Hasselbring
(Kiel University, Germany)

Preprint Available Video Info

Zenodo
Research. Shared

Experimental Data for: Exploring Software Cities through Virtual Reality

GitHub
Search GitHub

GitHub

Live trace visualization for large software landscapes

Modeling Polyp Activity of *Paragorgia arborea* Using Supervised Learning

Arne Johanson, Sascha Flögel, Wolf-Christian Dullo, Peter Linke, Wilhelm Hasselbring

- Software Engineering Group, Kiel University, Germany
- GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany

[Fig. 7. Degree of extension of individual polyps according to observations as well as to models with one, two, and six features. For an interactive illustration of this figure follow the link provided at: https://hub.com/ajohanson/paragorgia-arborea-activity.]

Publishing:
- **Paper:** http://dx.doi.org/10.1016/j.ecoinf.2017.02.007/
- **Code:** https://github.com/a-johanson/oceantea [Johanson et al. 2016]
- **Software service with data:** http://maui.se.informatik.uni-kiel.de:9090/ [Johanson et al. 2017]
**Note:**

**Computationally Replicable.** The experimental results of this paper were replicated by a SIGMOD Review Committee and were found to support the central results reported in the paper. Details of the review process are found [here](#).

**Authors:**

Ali Hadian
Sadegh Nobili
Behrooz Minouti
Qiang Qu

**Published in:**

- Proceeding of SIGMOD '16
  - Pages 182-193
  - San Francisco
  - ACM New York
  - [table of contents](#)

**Artifacts Review and Badging:**

A variety of research communities have embraced the goal of reproducibility in experimental science. [more information]

**Artifacts Evaluated – Functional**

The artifacts associated with the research are found to be documented, consistent, complete, exercisable, and include appropriate evidence of verification and validation.

**Artifacts Evaluated – Reusable**

The artifacts associated with the paper are of a quality that significantly exceeds minimal functionality.

**Artifacts Available**

Author-created artifacts relevant to this paper have been placed on a publically accessible archival repository.

**Results Replicated**

The main results of the paper have been obtained in a subsequent study by a person or team other than the authors, using, in part, artifacts provided by the author.

**Results Reproduced**

The main results of the paper have been independently obtained in a subsequent study by a person or team other than the authors, without the use of author-supplied artifacts.
Impact of Artifact Evaluation

Fig. 1. Average citation counts of AE and non-AE papers for conferences that used AE in 2013 to 2016 (conferences: VISSOFT, PPoPP, POPL, PLDI, PACT, OOPSLA, ISSTA, FSE, ECRTS, ECOOP, CGO, CAV).

Popper – Practical Falsifiable Research
http://falsifiable.us/
Förderangebote der DFG

Informations für die Wissenschaft Nr. 71 | 2. November 2016

Nachhaltigkeit von Forschungssoftware

DFG ermittelt Anträge zur Nutzbarmachung und Nachhaltigkeit von Forschungssoftware

Quelle: Dr. Matthias Katerbow (DFG LIS)
Summary

Key question:

• How to achieve **repeatability, extensibility** and **sustainability** of ocean research software?

Solution approach:

• Digital Ocean could invent and introduce new digital research infrastructures as supporting structures (mainly software) that are themselves **sustainable**.

Sustainable Development Goal 9

— “Build resilient **infrastructure**, promote inclusive and sustainable industrialization and foster innovation. [...]”
References


