Nachhaltige Softwareentwicklung für die Digitalisierung der Wissenschaft

Prof. Dr. Wilhelm (Willi) Hasselbring
Software Engineering Group
http://se.informatik.uni-kiel.de/

06. Juni 2018
Thema hier: 
Digitalisierung der Wissenschaft und Reproduzierbarkeit

https://doi.org/10.1109/MCSE.2018.108162940

[Johanson & Hasselbring 2018]
Gene name errors are widespread in the scientific literature

Mark Ziemann¹, Yotam Eren¹,² and Assam El-Osta¹,³*

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

Percentage of papers with gene lists affected

- Nature
- Genes Dev
- Genome Res
- Genome Biol
- Nat Genet
- Nucleic Acids Res
- BMC Genomics
- Overall Average
- RNA
- PLoS CompBiol
- PLoS Biology
- PLoS One
- Hum Mol Genet
- Science
- BMC Bioinformatics
- Bioinformatics
- Genome Biol Evol
- DNA Res
- Mol Biol Evol
A Challenge for Arne’s PhD research

Marine Biology Research
Publication details, including instructions for authors and subscription information: http://www.tandfonline.com/loi/smar20

Estimating the horizontal and temporal overlap of pelagic fish distribution in the Norwegian Sea using individual-based modelling
Kjell Rong Utne & Geir Huse

* Institute of Marine Research, Bergen, Norway

http://dx.doi.org/10.1080/17451000.2011.639781

• Utne & Huse provide an abstract (in part mathematical) description of their individual-based model, but:
  – We cannot reconstruct the implementation from the provided information
  – Sources for calibration data are named (some are unpublished) but again we cannot reconstruct the specific input data and parameters used.
• Without releasing the source code and the input/configuration data of the model, reproducibility of the results is hard or even impossible.
Reproducible Research in Computational Science

Roger D. Peng

“Replication is the ultimate standard by which scientific claims are judged.”
“Science advances faster when we can build on existing results, and when new ideas can easily be measured against the state of the art.”

At least Repeatability, not necessarily reproducibility

Several ACM SIGMOD, SIGPLAN, and SIGSOFT conferences have initiated artifact evaluation processes.
Artifact 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.
Artifact Evaluation Track

Chairs: Wilhelm Hasselbring (Kiel University) & Petr Tuma (Charles University)

Some numbers for ICPE 2018:

• 59 submitted full research papers
• 14 accepted full research papers
• 6 submitted artifacts
• 2 accepted artifacts, evaluated as functional
• 0 accepted artifacts, evaluated as reusable
Is it worth making the effort?

“If I have seen further, it is by standing on the shoulders of giants.”

- Isaac Newton

“Science advances faster when we can build on existing results, and when new ideas can easily be measured against the state of the art.”

[Krishnamurthi & Vitek 2015]
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).

[Childers & Chrysanthis 2017]
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

GitHub
Search GitHub

ExplorViz
Live trace visualization for large software landscapes
http://www.explorviz.net

[Fittkau et al. 2013, 2015a-e, 2017]
Cloud-Based Platform for Repeatable Ocean Observation Data Processing

OceanTEA

[Image of computer interface]

[Johanson et al. 2016a]
4D Spatial Analysis with OceanTEA
Machine Learning on Ocean Observation Data with OceanTEA

- Paper: http://dx.doi.org/10.1016/j.ecoinf.2017.02.007
- Source code: https://github.com/a-johanson/oceantea
- Software service with data: http://maui.se.informatik.uni-kiel.de:9090/
  (URL will change, refer to the GitHub repository for updates)

Modeling polyp activity of Paragorgia arborea using supervised learning

Arne N. Johanson, Suscha Flögel, Wolf-Christian Duto, Peter Linke, Wilhelm Hasselbring

[Fig. 7] Degree of expansion of coral 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://github.com/a-johanson/paragorgia-arborea-activity.
GeRDI
Generic Research Data Infrastructure

http://www.gerdi-project.de/
[Tavares de Sousa et al. 2018]
Analytic Reproducibility vs. Full Replication

Research Hypothesis

Scientific Experiment

Primary Data

Computer-Based Analysis

Analyzed Data

Scientific Publication

06.06.2018 W. Hasselbring
Ausblick: Digital Ocean 2018
https://digitalewochekiel.de/
References


References


