Good Scientific Data Management Practice

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

ISOS, March 22th, 2017
Agenda

• What’s the problem / challenge that I’m talking about?
  – And what I’m not talking about today …
• What’s the current state?
  – And what you could already do …
• What to expect in the future?
  – And what you could do …
Scientific misconduct and other challenges

• There exist several examples of scientific misconduct, such as the case Jan Hendrik Schön.
  – I skip that part in my presentation.

• However, there are also other challenges to obey the rules of good scientific practice,
  – that are not scientific misconduct.

• Let’s take a look at an example from the work of one of my former Ph.D. students...
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 \(^a\) & Geir Huse \(^a\)

\(^a\) 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  - Sources for calibration data are named (some are unpublished) but again we cannot reconstruct the specific input data and parameters used.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 \textbf{and} the input/configuration data of the model, \textit{reproducibility} of the results is hard or even impossible.
Reproducibility to Rectify Errors

PNAS 2016 vol. 113 no. 28 7900–7905, DOI: 10.1073/pnas.1602413113

Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund\textsuperscript{a,b,c,1}, Thomas E. Nichols\textsuperscript{d,e}, and Hans Knutsson\textsuperscript{a,c}

Significance

Functional MRI (fMRI) is 25 years old, yet surprisingly its most common statistical methods have not been validated using real data. Here, we used resting-state fMRI data from 499 healthy controls to conduct 3 million task group analyses. Using this null data with different experimental designs, we estimate the incidence of significant results. In theory, we should find 5\% false positives (for a significance threshold of 5\%), but instead we found that the most common software packages for fMRI analysis (SPM, FSL, AFNI) can result in false-positive rates of up to 70\%. These results question the validity of some 40,000 fMRI studies and may have a large impact on the interpretation of neuroimaging results.
Gene name errors are widespread in the scientific literature

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

Abstract

The spreadsheet software \textit{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
Recommendation 7 (of 16):

- Primary data as the basis for publications shall be securely stored for ten years in a durable form in the institution of their origin.
- Experiments and numerical calculations can only be repeated if all important steps are **reproducible**. For this purpose, they must be recorded.

(Source: http://doi.org/10.1002/9783527679188.oth1)

“If I have seen further it is by standing on the shoulders of giants.”
Isaac Newton, 1676
Reproducible Research in Computational Science

Roger D. Peng

“Replication is the ultimate standard by which scientific claims are judged.”

Reproducibility Spectrum

Publication only

Publication +

Code

Code and data

Linked and executable code and data

Full replication

Not reproducible

Gold standard
So, what’s the problem / challenge that I’m talking about?

• For good scientific practice, it is important that research results may be
  – properly checked by reviewers and
  – possibly repeated and extended by other researchers.

• This is of particular interest for “digital science” i.e. for in silico experiments

• How can Software Systems and Services Contribute?
What I’m not talking about?

Software and services for detecting plagiarism, such as

http://plagiarism-detector.com/

https://www.plagaware.com/
What I’m not talking about?

• Establishing Software Engineering best practices in computational science, such as
  – Version and configuration management
    [Ploski et al. 2007]
  – Quality management
    [Waller et al. 2015, Hasselbring & Steinacker 2017]
  – Parallel and distributed programming

• To learn about such topics, you may attend my regular lectures (BSc, MSc) in Computer Science
Agenda

• What’s the problem / challenge that I’m talking about?
  – And what I’m not talking about today ...

• What’s the current state?
  – And what you could already do ...

• What to expect in the future?
  – And what you could do ...
Research Workflows

1. Define the research question
2. Gather information and resources (observe)
3. Form hypothesis
4. Perform experiment and collect data
5. Analyze data
6. Interpret data and draw conclusions that serve as a starting point for new hypothesis
7. Publish results
8. Retest (frequently done by other scientists)

Crawford S, Stucki L (1990), "Peer review and the changing research record", J Am Soc Info Science", vol. 41, pp 223-228
Data Repositories (Services): Examples
[Registry: http://www.re3data.org/]

PANGAEA

Data Publisher for Earth & Environmental Science
https://www.pangaea.de/

World Data Center
for Climate
Hamburg

https://www.dkrz.de/daten/wdcc/

Zenodo

http://zenodo.org/

22.03.2017 W. Hasselbring
Virtuelle Forschungsumgebung in Kiel

Forschungsdatenmanagement

Forschungsdaten sind die Daten, die in einer virtuellen Forschungsumgebung vorgehalten werden, und sind die Grundlage wissenschaftlichen Arbeitens.

Allgemeine Informationen zum Forschungsdatenmanagement und Informationen mit Bezug zum Forschungsstandort Kiel sind auf den folgenden Seiten zusammengestellt.

Aktivitäten

lokale Angebote, und Informationen zu Vernetzungen und Kooperationen über den Standort Kiel hinaus sind hier gelistet.

Kolloquium Forschungsdatenmanagement

Das Kolloquium Forschungsdatenmanagement ist eine regelmäßig stattfindende Veranstaltungsreihe, in der die Möglichkeit zum Austausch über den Umgang mit Forschungsdaten über die Grenzen der verschiedenen Disziplinen an der CAU und ein Forum zur Diskussion aktueller Themen geboten werden. Hier weiterlesen...

Umfrage zum Umgang mit Forschungsdaten

Zur Ermittlung der vielfältigen Anforderungen der verschiedenen Fachdisziplinen wurde eine Umfrage zum Thema 'Umgang mit Forschungsdaten am Forschungsstandort Kiel' durchgeführt. Hier weiterlesen...

Kontext und Information

Nationale, europäische und internationale Rahmenbedingungen und damit zusammenhängende Informationen zum Thema Forschungsdaten sind hier gelistet.

Plattform Data Management


Neben der Kiel Data Management Infrastructure (KDMI) für Meereswissenschaften bilden die DFG-Projekte 'PubFlow' der AG Software Engineering und 'VFÜ@Kiel' des Servicezentrums Forschung und Innovation und Rechenzentrums die wesentliche Grundlage für flexibles Forschungsdatenmanagement an der CAU.

Ziele

Aufgaben

Leitlinie zum Umgang mit Forschungsdaten

AG Forschungsdatenmanagement

Kolloquium Forschungsdatenmanagement

Projekte

http://www.uni-kiel.de/vfu/de/forschungsdatenmanagement
Kiel Data Management Portal

Refer to the “hands on” seminars

https://portal.geomar.de/
Kiel Data Management Infrastructure – OSIS: Ocean Science Information System

Source: Kiel Data Management Team, https://portal.geomar.de/kdmi
OceanRep link to OSIS data, and back
OceanRep link to Pangaea

It is the economy, stupid! Projecting the fate of fish populations using ecological-economic modeling


Kielprints is a similar service for Kiel at large: http://eprints.uni-kiel.de
Agenda

• What’s the problem / challenge that I’m talking about?
  – And what I’m not talking about today ...

• What’s the current state?
  – And what you could already do ...

• What to expect in the future?
  – And what you could do ...
Toward Publication Workflows

Funded:
Software Engineering Group, University Kiel

Associated:
• Excellence cluster “Future Ocean”
• Data and computing center of GEOMAR
• Library of GEOMAR
• Computing center of University Kiel
• Library of University Kiel
• ZBW

German National Library of Economics - Leibniz Information Centre for Economics

http://www.pubflow.uni-kiel.de/ [Brauer & Hasselbring 2013]
Data and Paper Flow (in Ocean Science)
CTD Workflow
Need to save data + processing (not yet addressed in PubFlow)

Algorithms + Data Structures = Programs

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.”
Viewpoint
The Real Software Crisis: Repeatability as a Core Value

Sharing experiences running artifact evaluation committees for five major conferences.

“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

Experimental Data for: Exploring Software Cities through Virtual Reality

GitHub
Search GitHub

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

[Fittkau et al. 2013, 2015a-c, 2016]
Cloud-Based Platform for Repeatable Ocean Observation Data Processing

OceanTEA

https://github.com/a-johanson/oceantea [Johanson et al. 2016a]

22.03.2017 W. Hasselbring
Modeling Polyp Activity of Paragorgia arborea Using Supervised Learning

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

Abstract—While the distribution patterns of cold-water corals, such as Paragorgia arborea, have received increasing attention in recent studies, little is known about their in situ activity patterns. In this paper, we examine polyp activity in P. arborea using machine learning techniques to analyze high-resolution time-series data and photographs obtained from an autonomous lander cluster deployed in the Stýrmundur, Norway. An interactive illustration of the models derived in this paper is provided online as supplementary material.

We find that the best predictor of the degree of extension of the coral polyps is current direction with a lag of three hours. Other variables that are not directly associated with water currents, such as temperature and salinity, offer much less information concerning polyp activity. Interestingly, the degree of polyp extension can be predicted more reliably by sampling the laminar flows in the water column above the measurement site than by sampling the more turbulent flows in the direct vicinity of the corals.

Our results show that the activity patterns of the P. arborea polyps are governed by the strong tidal current regime of the Stýrmundur. It appears that P. arborea does not react to shorter changes in the ambient current regime but instead adjusts its behavior in accordance with the large-scale pattern of the tidal cycle itself in order to optimize nutrient uptake.

1 Introduction

Cold-water corals (CWCs) such as Paragorgia arborea and Lophelia pertusa can be found on continental shelves, slopes, and seamounts all over the world. Like tropical coral reefs, which inhabit shallower and warmer waters, CWC reefs are associated with high biodiversity as they provide habitat to many other species (Roberts and Cairns 2014; Roberts et al. 2006).
# GeRDI

**Generic Research Data Infrastructure**

<table>
<thead>
<tr>
<th>Economics</th>
<th>Life science, Humanities</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOEP</td>
<td>CBG</td>
</tr>
<tr>
<td></td>
<td>Max Planck Institute</td>
</tr>
<tr>
<td></td>
<td>of Molecular Cell Biology</td>
</tr>
<tr>
<td></td>
<td>and Genetics</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Marine science</th>
<th>Environmental science</th>
</tr>
</thead>
<tbody>
<tr>
<td>KIEL Marine</td>
<td>ClimEx</td>
</tr>
<tr>
<td>Science</td>
<td></td>
</tr>
<tr>
<td>Connected</td>
<td></td>
</tr>
<tr>
<td>Research</td>
<td></td>
</tr>
</tbody>
</table>

- [Gerdi-project.de](http://www.gerdi-project.de/)

---

22.03.2017  W. Hasselbring

---

**ZBW**

Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre for Economics

**Technische Universität Dresden**

**Christian-Albrechts-Universität zu Kiel**

**DFG**
Envisioned GeRDI Architecture
What about social networks?

Specific social networks for academics exist such as ResearchGate (http://www.researchgate.net) or Mendeley (http://www.mendeley.com/).
Policies and Incentives

• Funding agencies, such as the DFG, require strategies for research data management
  – Institutional data policies and infrastructures may help
  – “Modular” data management policy for Kiel Marine Sciences may already be reused

• Published data and code may be listed in CVs

• Cost benefit analysis of the DRYAD repository
  – Papers with published data receive higher citation counts:
    • Piwowar, Vision, Whitlock: “Data archiving is a good investment”, Nature 473(285), 2011 http://dx.doi.org/10.1038/473285a
Summary

• If you are only interested in getting a Ph.D.,
  – this talk was not really of interest to you, sorry.
• If you are (also) interested in scientific impact, publish
  – research papers,
  – research data,
  – documented code, and
  – do networking with related stakeholders.
• Software systems and services may help
• Outlook:
  – “Digital Ocean” in “Future Ocean Sustainability”

You may find these slides at: http://eprints.uni-kiel.de/37072
References


References


