Good Scientific Data Management Practice

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

ISOS, March 22th, 2017

future ocean
KIEL MARINE SCIENCES

KIEL MARINE SCIENCE
CONNECTED RESEARCH

C A U
Christian-Albrechts-Universität zu Kiel

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

Helmholtz Ocean System Science and Technology

Halifax MARINE RESEARCH Institute · DALHOUSIE UNIVERSITY Inspiring Minds · C A U Christian-Albrechts-Universität zu Kiel · GEOMAR

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
Version of record first published: 25 Apr 2012.
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  - Sources for calibration data are named (some are unpublished) but again we cannot reconstruct the specific input data and parameters used.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.

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[a,b,c,1], Thomas E. Nichols[d,e], and Hans Knutsson[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.

Genome Biology 2016 17:177 **DOI:** 10.1186/s13059-016-1044-7

Gene name errors are widespread in the scientific literature

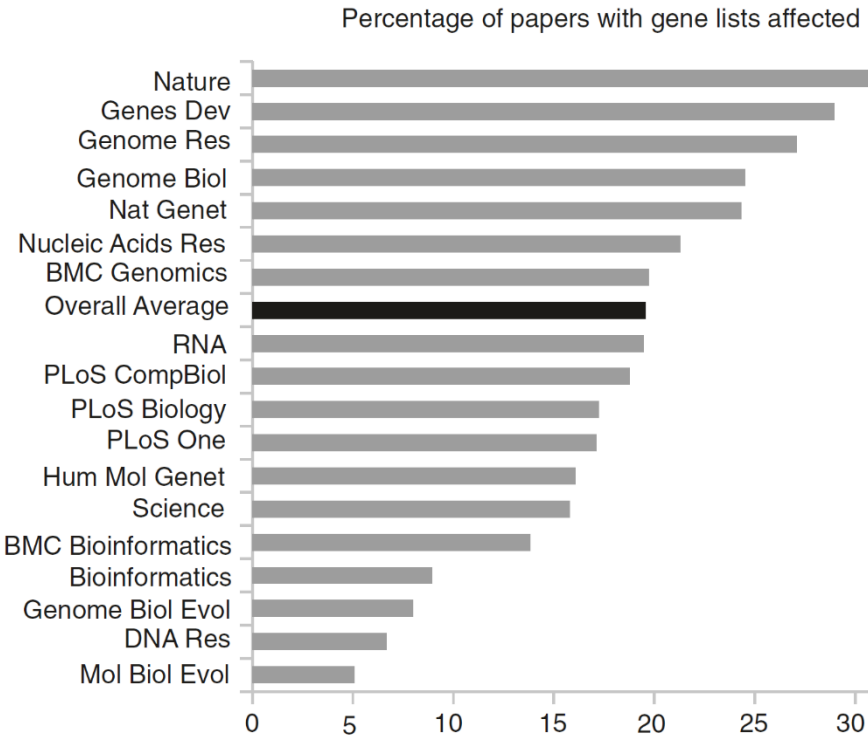
Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

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



Recommendations of the Commission on Professional Self Regulation in Science

Proposals for Safeguarding Good Scientific Practice
January, 1998

DFG

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

Science 2 December 2011:
Vol. 334 no. 6060 pp. 1226-1227
DOI: 10.1126/science.1213847

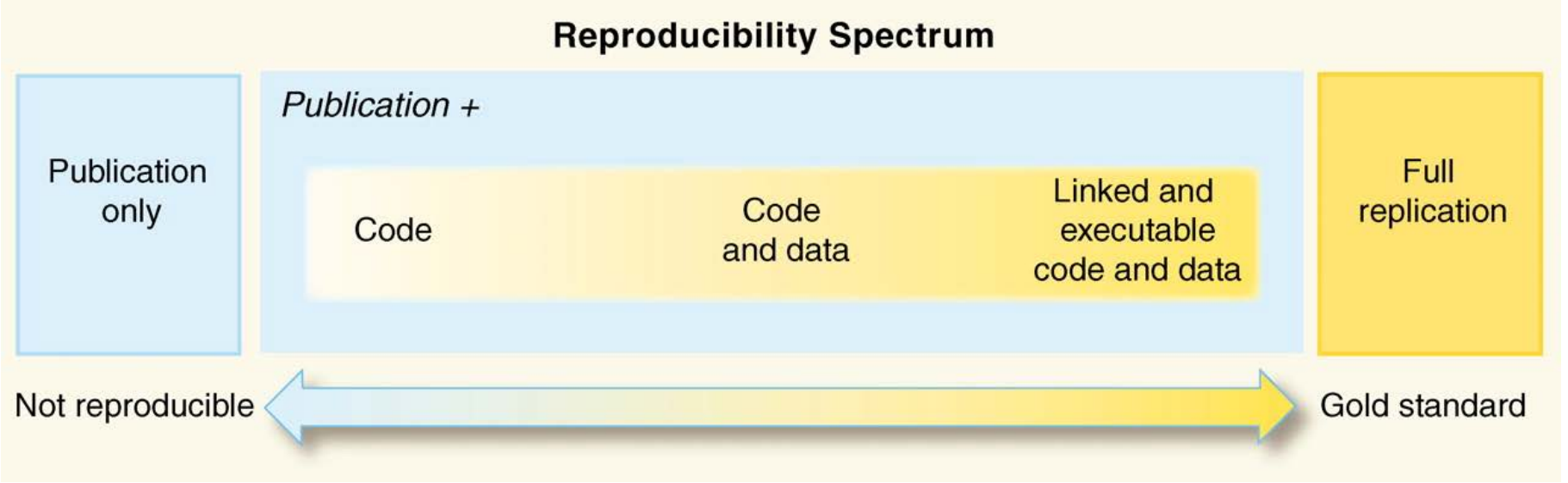
Science

PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

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



So, what's the problem / challenge that I'm talking about?

David Zinn, Introduction to the Responsible Conduct of Research

- 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/

PlagAware
We discover Plagiarism.

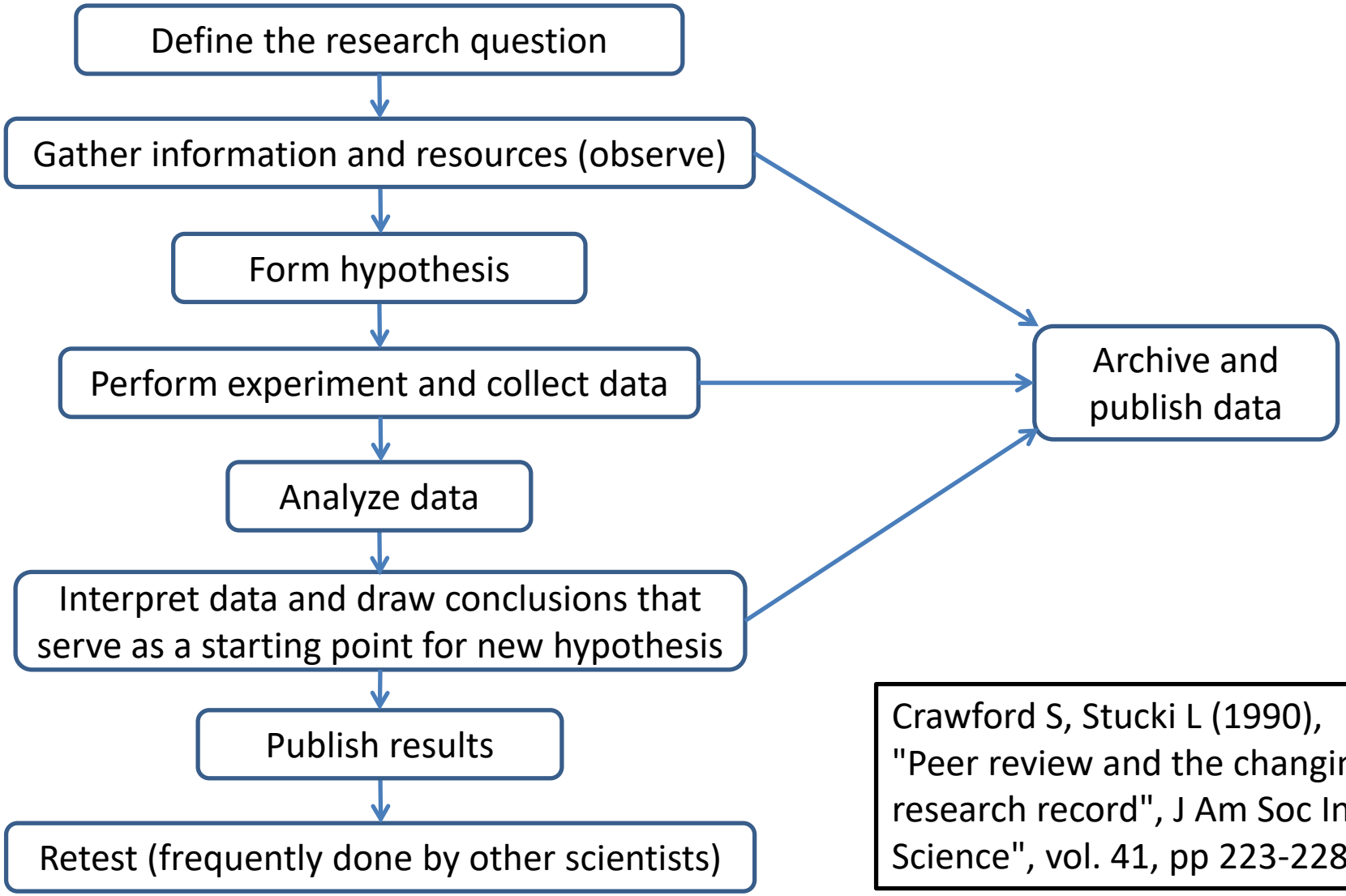
What I'm not talking about?

- 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]
  - **Software architecture design and modeling** [Hasselbring 2002, Hasselbring 2006, Reussner & Hasselbring 2008]
  - **Domain-specific programming languages** [Johanson & Hasselbring 2016, Johanson et al. 2016b, Johanson et al. 2017a]
  - **Parallel and distributed programming** [Hasselbring 1994, Hasselbring 2000, Wulf et al. 2016]
- To learn about such topics, you may attend my regular lectures (BSc, MSc) in Computer Science

Agenda

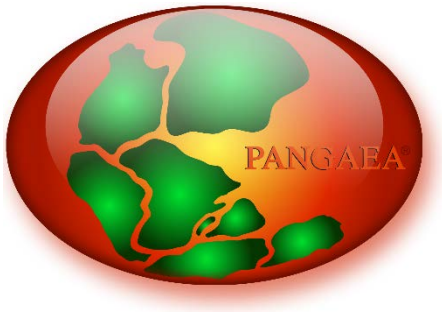
- What's the problem / challenge that I'm talking about?
  - And what I'm not talking about today …
- **What's the current state?**
  - **And what you could already do …**
- What to expect in the future?
  - And what you could do …

Research Workflows



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/]



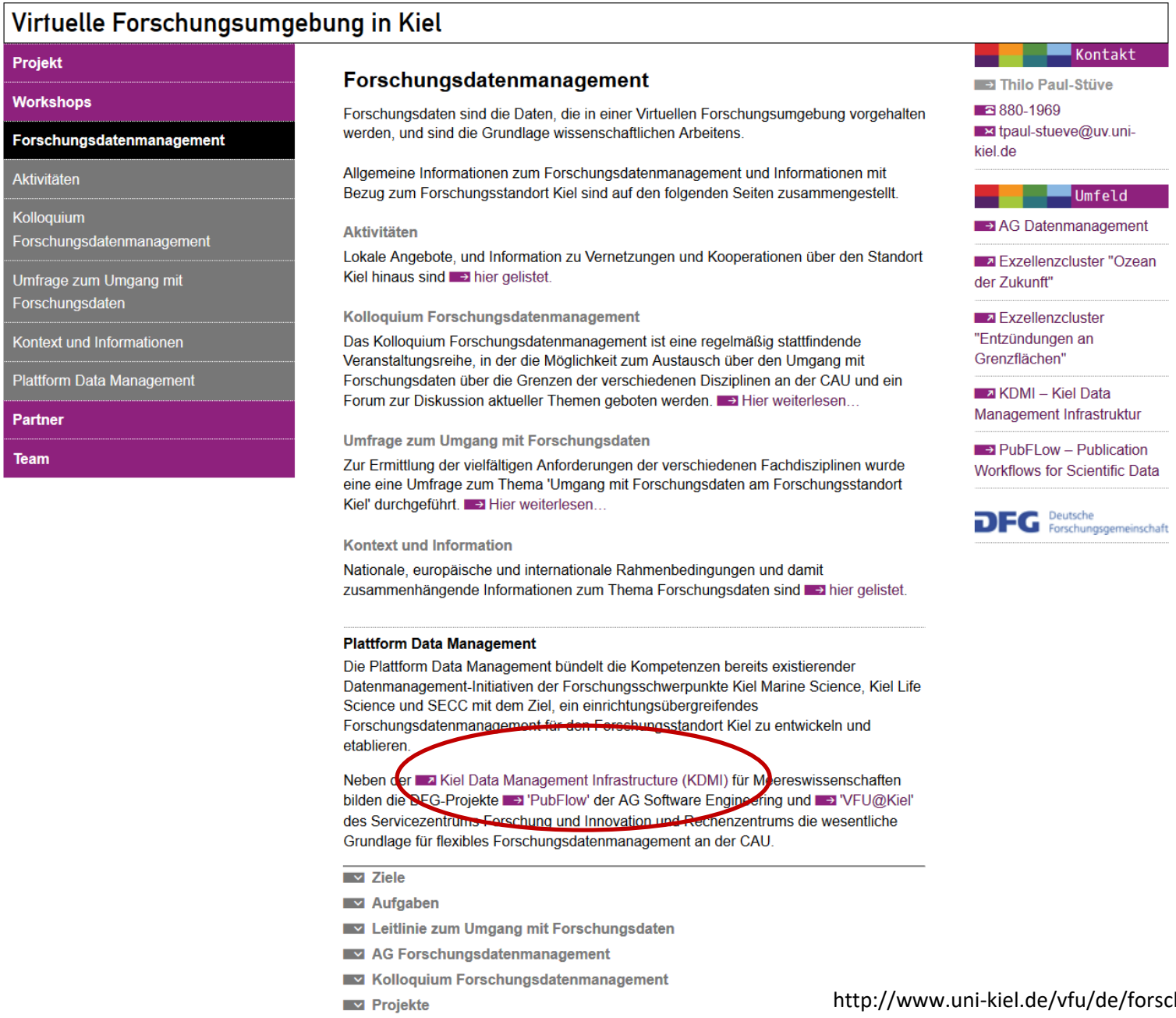
PANGAEA®

Data Publisher for Earth & Environmental Science

https://www.pangaea.de/

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

http://zenodo.org/

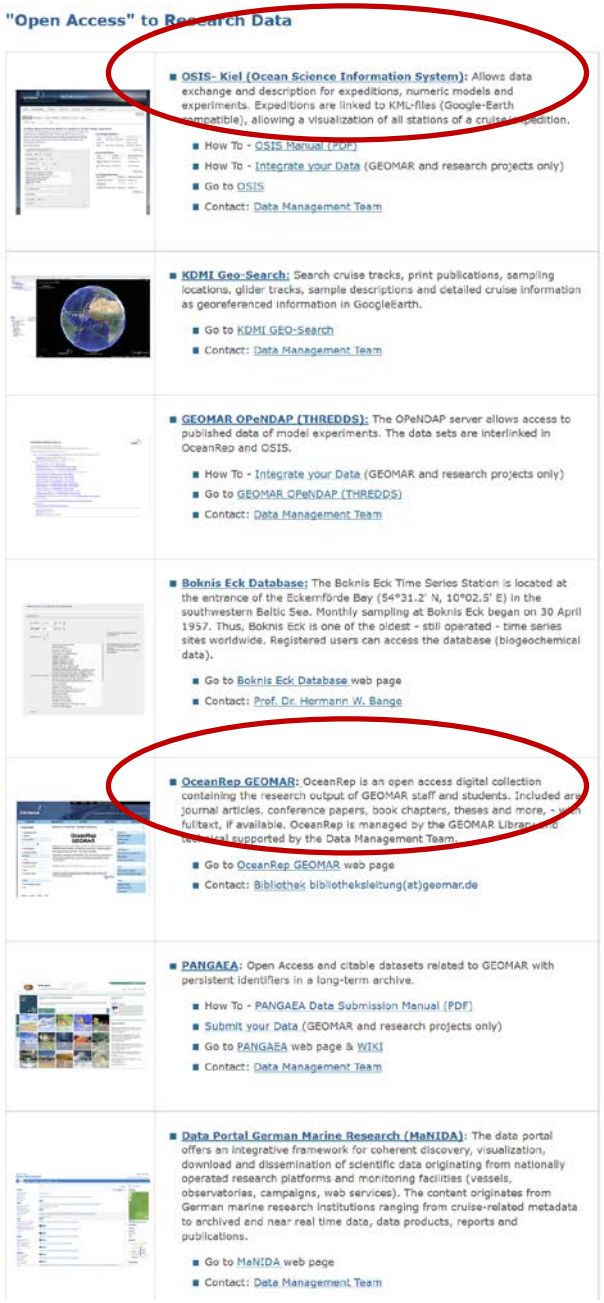


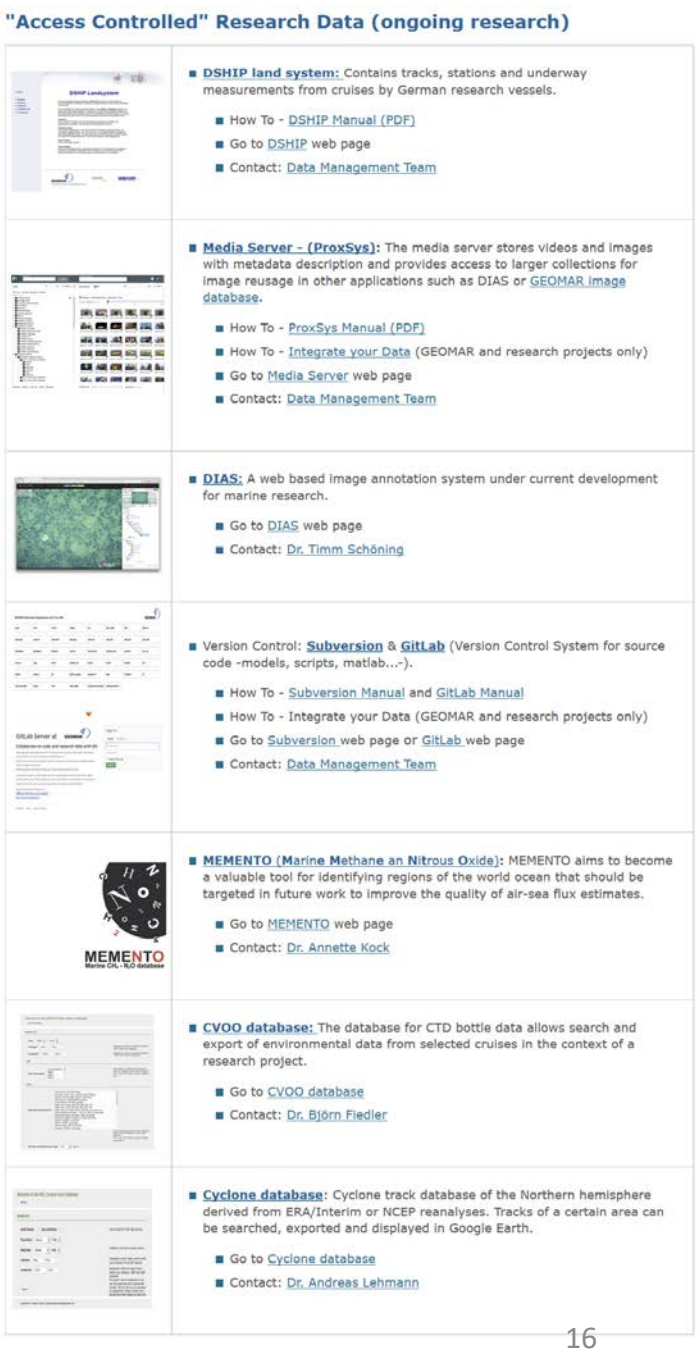
http://www.uni-kiel.de/vfu/de/forschungsdatenmanagement

Kiel Data Management Portal

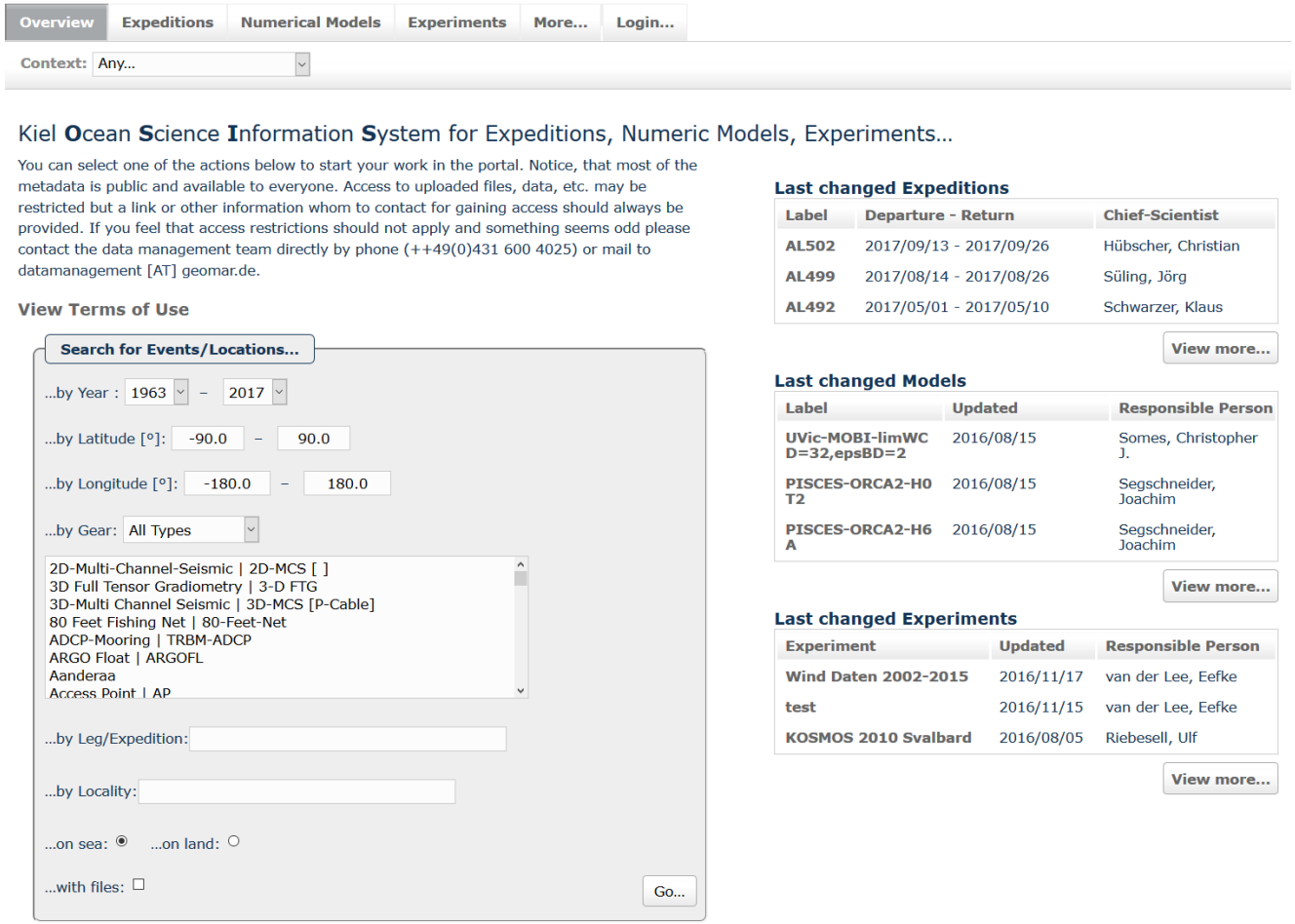
Refer to the “hands on” seminars

https://portal.geomar.de/



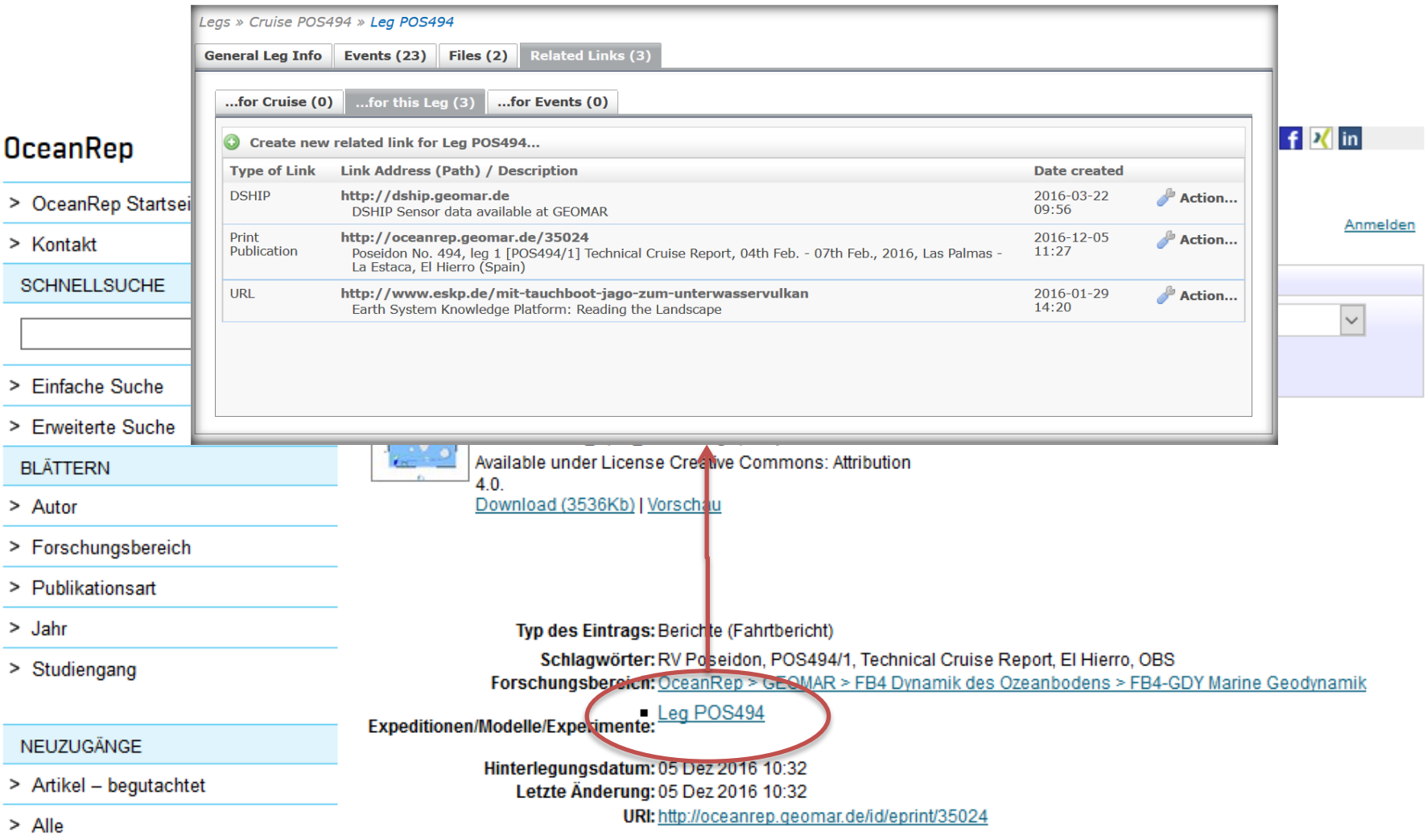


Kiel Data Management Infrastructure – OSIS: Ocean Science Information System

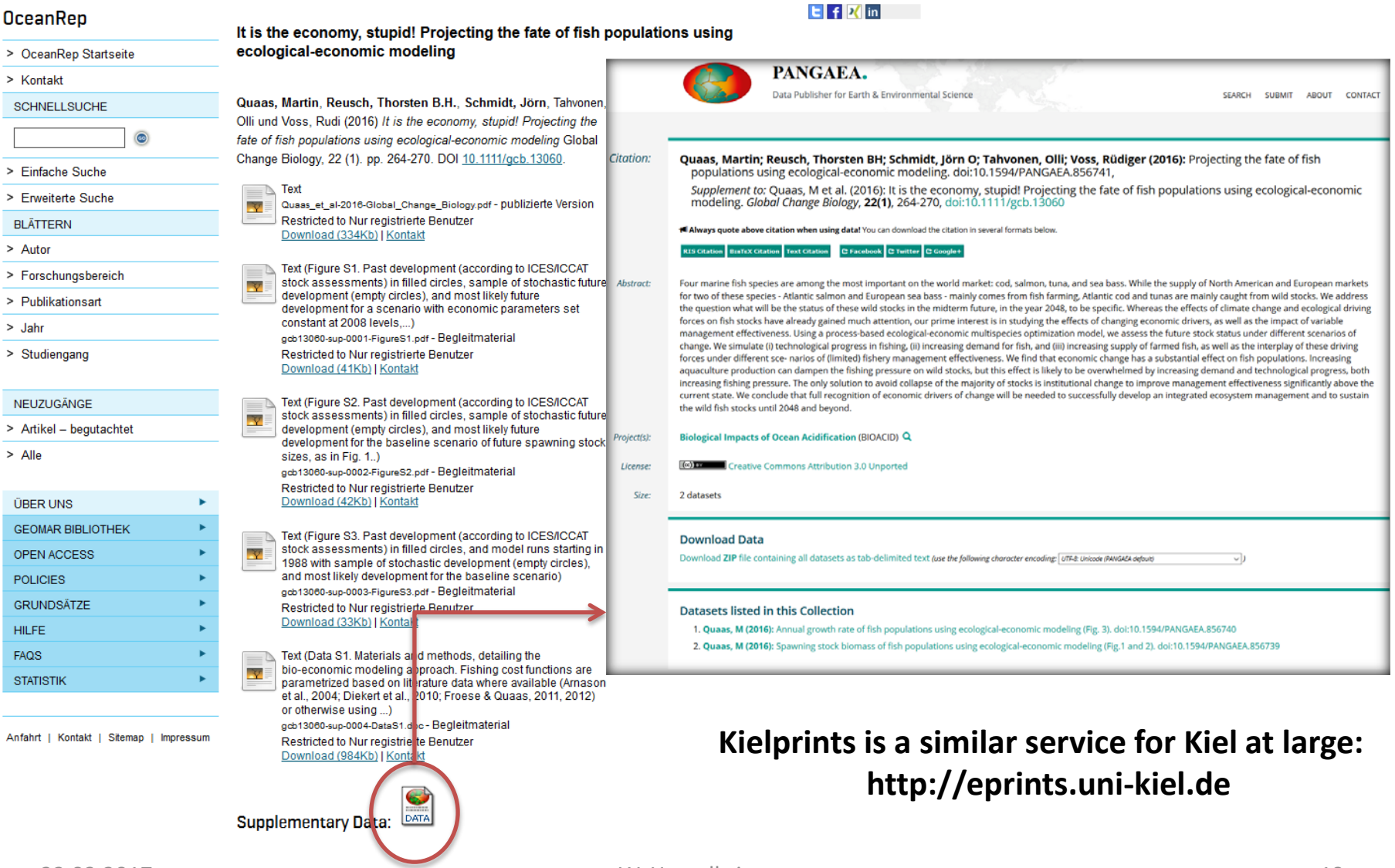


Source: Kiel Data Management Team, https://portal.geomar.de/kdmi

OceanRep link to OSIS data, and back



OceanRep link to Pangaea



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:

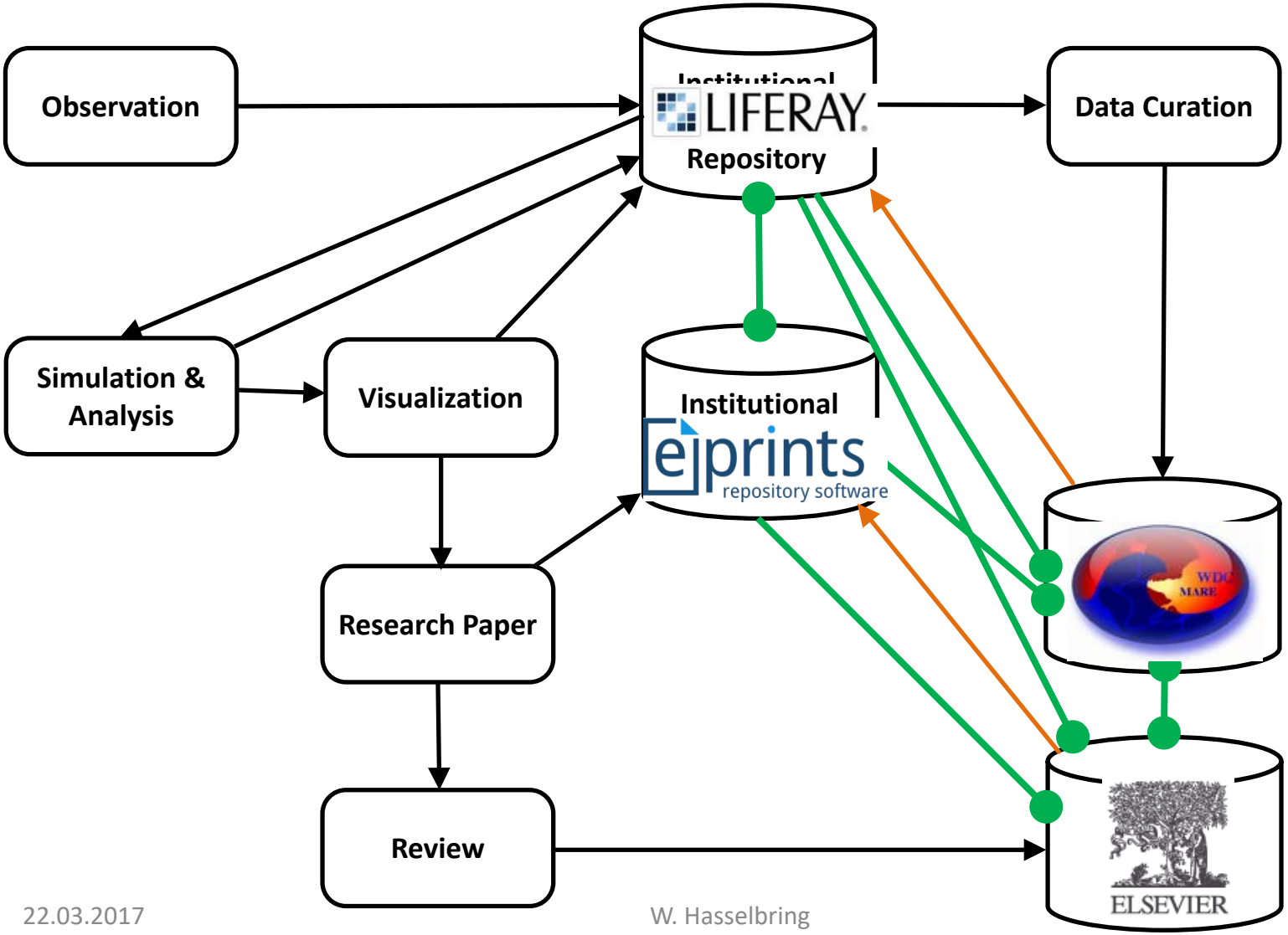
Software Engineering Group, University Kiel

Associated:

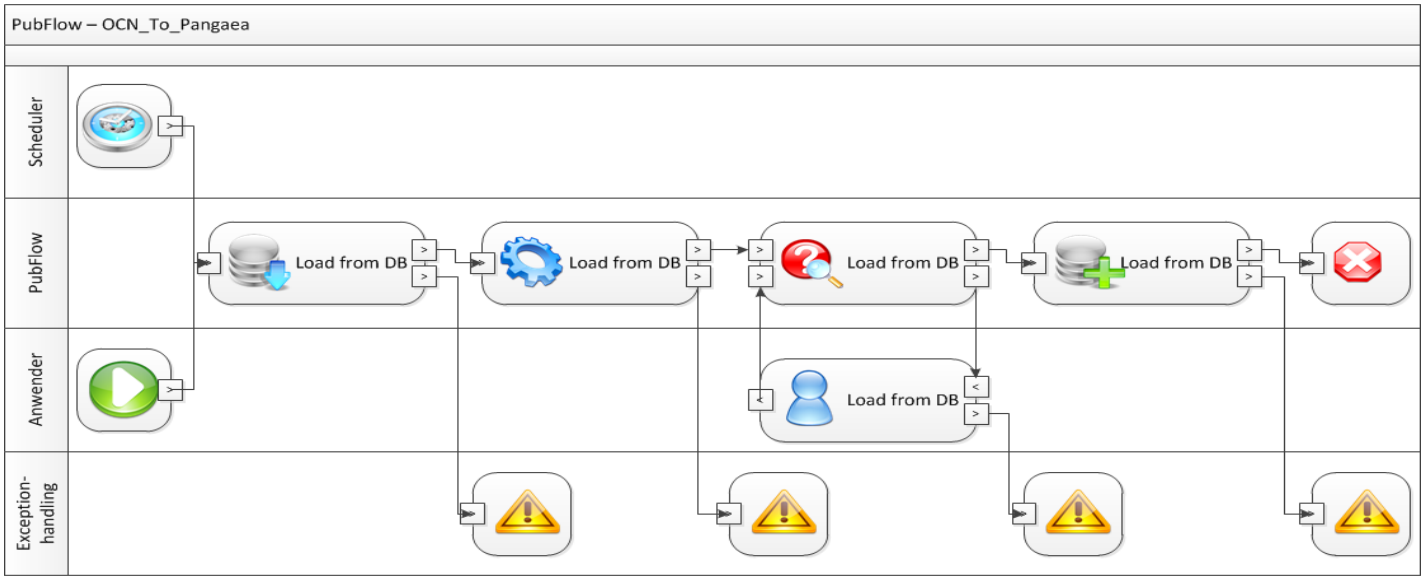
- Excellence cluster “Future Ocean”
- Data and computing center of GEOMAR
- **Library** of GEOMAR
- Computing center of University Kiel
- **Library** of University Kiel
- ZBW
  German National **Library** of Economics - Leibniz Information Centre for Economics

http://www.pubflow.uni-kiel.de/ [Brauer & Hasselbring 2013]

Data and Paper Flow (in Ocean Science)



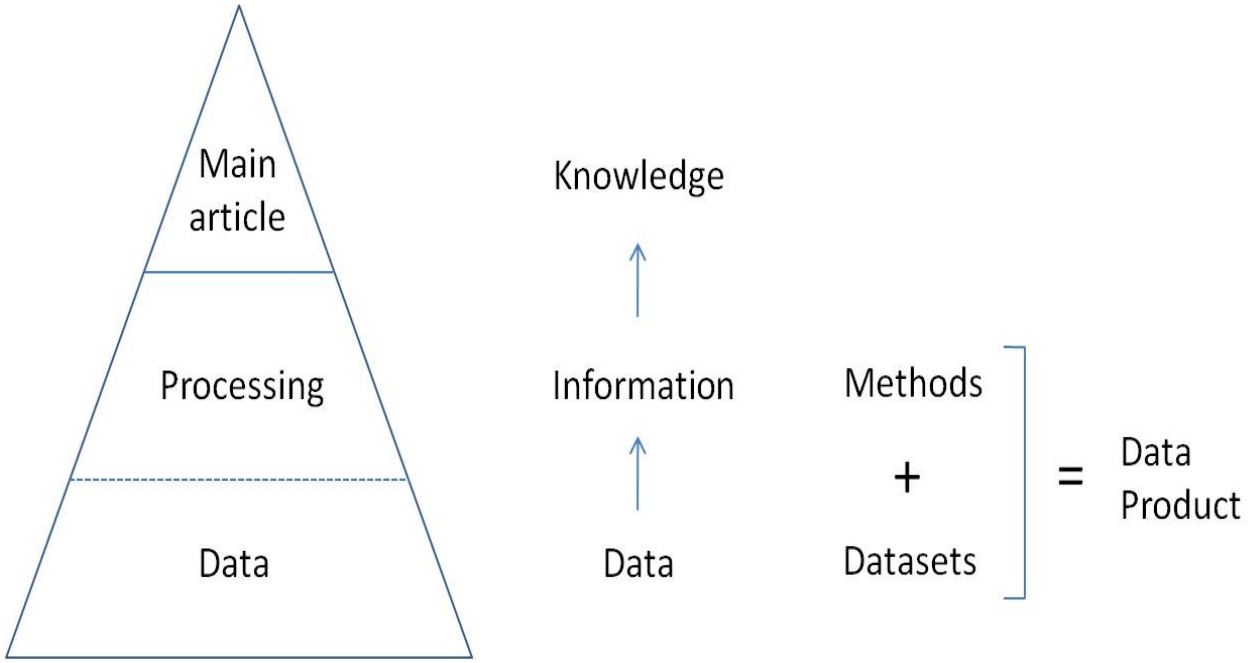
CTD Workflow



JBoss® jBPM

JIRA

Need to save data + processing (not yet addressed in PubFlow)



Algorithms + Data Structures = Programs

Source: Kunze, John A; Cruse, Patricia; Hu, Rachael; Abrams, Stephen; Hastings, Kirk; Mitchell, Catherine; et al. (2011). Practices, Trends, and Recommendations in Technical Appendix Usage for Selected Data-Intensive Disciplines. http://escholarship.org/uc/item/9jw4964t

PERSPECTIVE

nature

doi:10.1038/nature10836

The case for open computer programs

Darrel C. Ince[1], Leslie Hatton[2] & John Graham-Cumming[3]

- "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."

COMMUNICATIONS OF THE ACM | MARCH 2015 | VOL. 58 | NO. 3

DOI:10.1145/2658987

Shriram Krishnamurthi and Jan Vitek

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*

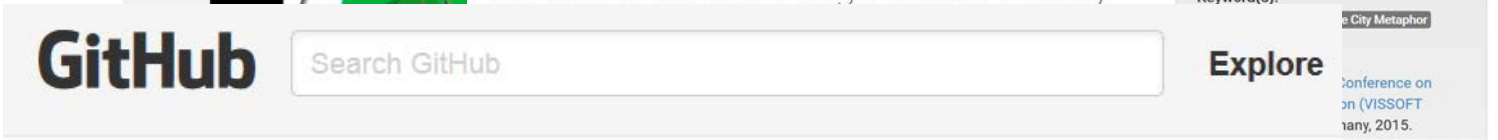
Several ACM SIGMOD, SIGPLAN, and SIGSOFT conferences have initiated artifact evaluation processes.

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

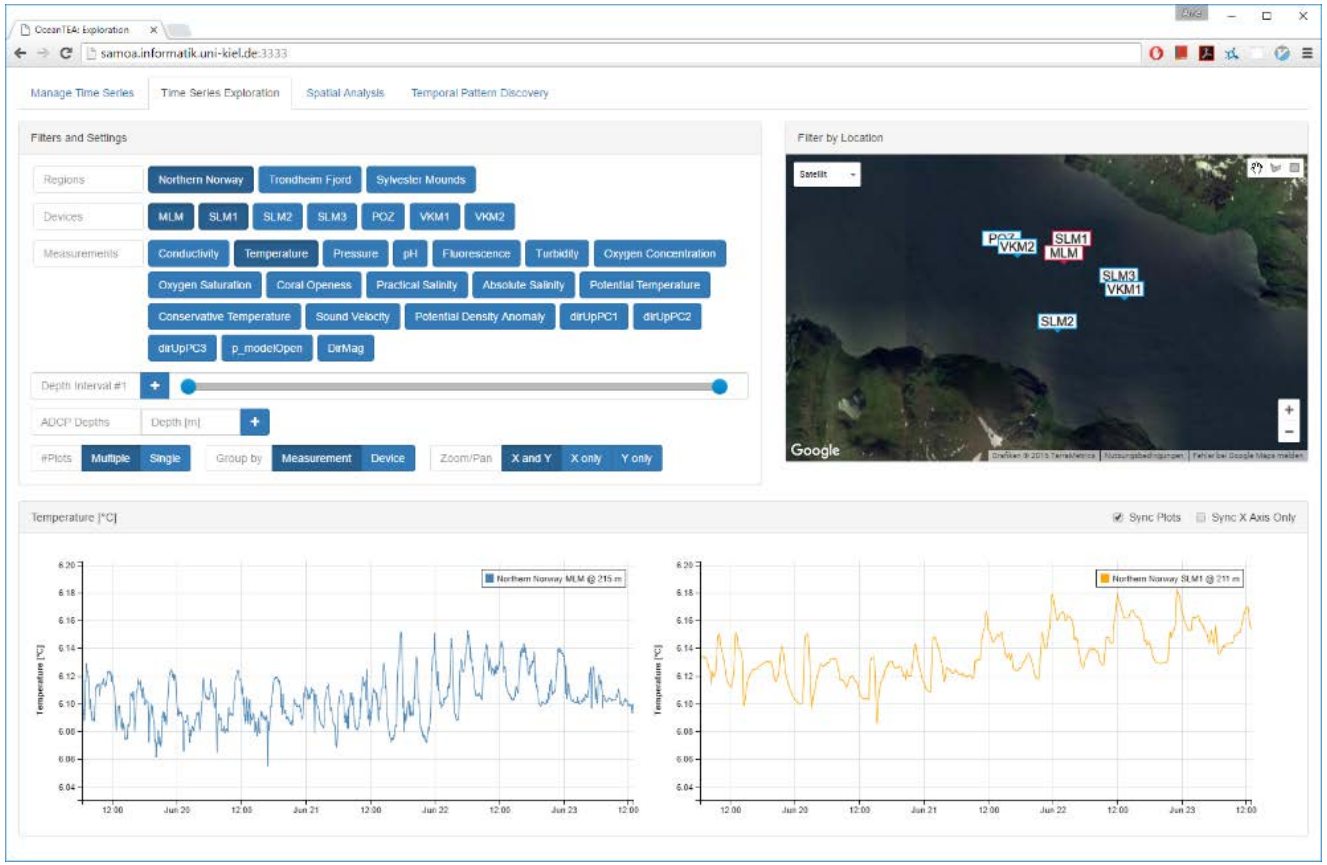


ExplorViz
Live trace visualization for large software landscapes
http://www.explorviz.net

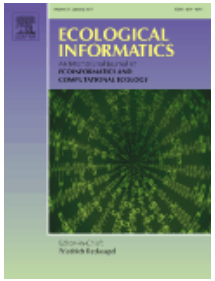
[Fittkau et al. 2013, 2015a-c, 2016]

Cloud-Based Platform for Repeatable Ocean Observation Data Processing

OceanTEA



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



Publishing:

- Paper: https://www.journals.elsevier.com/ecological-informatics/
- Code: https://github.com/a-johanson/oceantea
- Software service with data: http://maui.se.informatik.uni-kiel.de:9090/

Modeling Polyp Activity of *Paragorgia arborea* Using Supervised Learning

Arne Johanson,[a] Sascha Flögel,[b] Wolf-Christian Dullo,[b]
Peter Linke,[b] Wilhelm Hasselbring[a]

[a] Software Engineering Group, Kiel University, Germany
[b] GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany

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 Stjernsund, 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 Stjernsund. 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).

[Johanson et al. 2017b]

GeRDI
Generic Research Data Infrastructure

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

TECHNISCHE UNIVERSITÄT DRESDEN

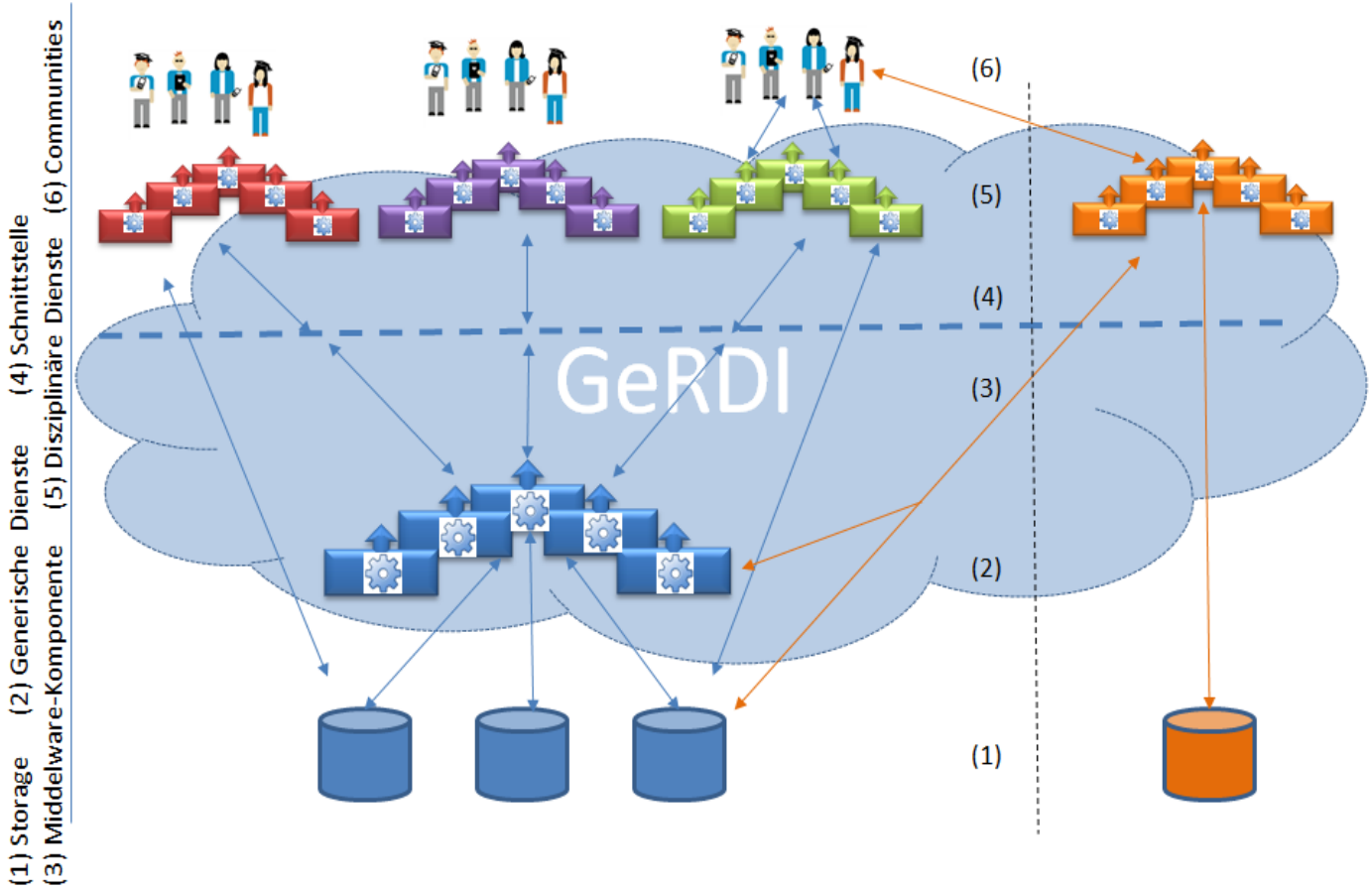
Economics	Life science, Humanities
SOEP	CBG Max Planck Institute of Molecular Cell Biology and Genetics; dh
Marine science	**Environmental science**
KIEL MARINE SCIENCE; PANGAEA	ClimEx

C A U
Christian-Albrechts-Universität zu Kiel

lrz

http://www.gerdi-project.de/

Envisioned GeRDI Architecture



What about social networks?

Social Networking Meets Software Development:

Perspectives from GitHub, MSDN, Stack Exchange, and TopCoder

Andrew Begel, Microsoft

Jan Bosch, Chalmers University of Technology

Margaret-Anne Storey, University of Victoria

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

IEEE Software 30(1): 26-28, 2013
Digital Object Identifier: 10.1109/MS.2013.13

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

[Brauer & Hasselbring 2013] P. C. Brauer, W. Hasselbring: “PubFlow: a scientific data publication framework for marine science”. In: International Conference on Marine Data and Information Systems (IMDIS 2013), 2013, Lucca, Italy.

[Fittkau et al. 2013] F. Fittkau, J. Waller, C. Wulf, W. Hasselbring: “Live Trace Visualization for Comprehending Large Software Landscapes: The ExplorViz Approach“, In: 1st IEEE International Working Conference on Software Visualization (VISSOFT 2013).

[Fittkau et al. 2015a] F. Fittkau, S. Roth, W. Hasselbring: “ExplorViz: Visual Runtime Behavior Analysis of Enterprise Application Landscapes“, In: 23rd European Conference on Information Systems (ECIS 2015).

[Fittkau et al. 2015b] F. Fittkau, A. Krause, W. Hasselbring: “Hierarchical Software Landscape Visualization for System Comprehension: A Controlled Experiment”. In: 3rd IEEE Working Conference on Software Visualization, 2015.

[Fittkau et al. 2015c] F. Fittkau, A. Krause, W. Hasselbring: “Exploring Software Cities in Virtual Reality”, In: 3rd IEEE Working Conference on Software Visualization, September 2015, Bremen, Germany.

[Fittkau et al. 2016] F. Fittkau, A. Krause, W. Hasselbring: “Software landscape and application visualization for system comprehension with ExplorViz”, In: Information and Software Technology. DOI 10.1016/j.infsof.2016.07.004

[Hasselbring 1994] W Hasselbring: “Prototyping Parallel Algorithms in a Set-Oriented Language, Verlag Dr. Kovac, 1994.

[Hasselbring 2000] W. Hasselbring: “Programming Languages and Systems for Prototyping Concurrent Applications”, In: ACM Computing Surveys, 31 (1). pp. 43-79. DOI 10.1145/349194.349199.

[Hasselbring 2002] W. Hasselbring: “Component-Based Software Engineering”, In: Handbook of Software Engineering and Knowledge Engineering. World Scientific Publishing, Singapore, pp. 289-305, 2002.

[Hasselbring 2006] W. Hasselbring: “Software-Architektur”. In: Informatik-Spektrum 29(1): 48-52, February 2006.

[Hasselbring & Steinacker 2017] W. Hasselbring, G. Steinacker: “Microservice Architectures for Scalability, Agility and Reliability in E-Commerce”, In: Proceedings of the IEEE International Conference on Software Architecture (ICSA 2017), April 2017, Gothenburg, Sweden.

[Heinrich et al. 2014] R. Heinrich, E. Schmieders, R. Jung, K. Rostami, A. Metzger, W. Hasselbring, R. Reussner, K. Pohl: “Integrating Run-Time Observations and Design Component Models for Cloud System Analysis“, In: 9th Workshop on Models@run.time 2014.

References

[Heinrich et al. 2015] R. Heinrich, R. Jung, E. Schmieders, A. Metzger, W. Hasselbring, R. Reussner, K. Pohl: “Architectural Run-Time Models for Operator-in-the-Loop Adaptation of Cloud Applications”, In: 9th IEEE Symposium on the Maintenance and Evolution of Service-Oriented Systems and Cloud-Based Environments (MESOCA 2015).

[Johanson et al. 2016a] A. Johanson, S. Flögel, C. Dullo, W. Hasselbring: “OceanTEA: Exploring Ocean-Derived Climate Data Using Microservices”. In: Sixth International Workshop on Climate Informatics (CI 2016), September 2016, Boulder, Colorado.

[Johanson et al. 2016b] A. Johanson, W. Hasselbring, A. Oschlies, B. Worm: “Evaluating Hierarchical Domain-Specific Languages for Computational Science: Applying the Sprat Approach to a Marine Ecosystem Model”. In: Software Engineering for Science. CRC Presspp. 175-200. DOI 10.1201/9781315368924-9.

[Johanson et al. 2017a] A. Johanson, A. Oschlies, W. Hasselbring, A. Worm: “SPRAT: A spatially-explicit marine ecosystem model based on population balance equations”, In: Ecological Modelling, DOI 10.1016/j.ecolmodel.2017.01.020.

[Johanson et al. 2017b] A. Johanson, S. Flögel, C. Dullo, P. Linke, W. Hasselbring: “Modeling Polyp Activity of Paragorgia arborea Using Supervised Learning”, In: Ecological Informatics, Elsevier, 2017.

[Johanson & Hasselbring 2016] A. Johanson, W. Hasselbring: “Effectiveness and efficiency of a domain-specific language for high-performance marine ecosystem simulation: a controlled experiment”, In: Empirical Software Engineering. DOI 10.1007/s10664-016-9483-z.

[Ploski et al. 2007] J. Ploski, W. Hasselbring, S. Schwierz, J. Rehwinkel: “Introducing Version Control to Database-Centric Applications in a Small Enterprise”, In: IEEE Software, 24 (1). pp. 38-44. DOI 10.1109/MS.2007.17.

[Reussner & Hasselbring 2008] R. Reussner, W. Hasselbriung: “Handbuch der Software-Architektur”. dpunkt.verlag, 2nd edition, 2008.

[Waller et al. 2015] J. Waller, N. Ehmke, W. Hasselbring: “Including Performance Benchmarks into Continuous Integration to Enable DevOps“, In: ACM SIGSOFT Software Engineering Notes, 40(2).

[Wulf et al. 2016] C. Wulf, C.C. Wiechmann, W. Hasselbring: “Increasing the Throughput of Pipe-and-Filter Architectures by Integrating the Task Farm Parallelization Pattern”, In: 18th International ACM SIGSOFT Symposium on Component-Based Software Engineering, April 2016, Venice, Italy.