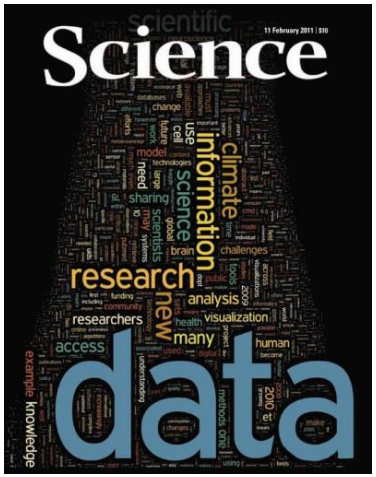
Data Management:
How can Software Systems and Services Contribute to Good Scientific Practice?

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

ISOS, December 3rd, 2013



C | A | U
Christian-Albrechts-Universität zu Kiel

future ocean
KIEL MARINE SCIENCES

My Background (short extract)

- CAU Kiel, Department of Computer Science, since 2008
  - Head, Software Engineering Group
  - Previously: University of Oldenburg (similar position 2000-2008), Tilburg (NL), Dortmund, Essen, Braunschweig.
- Competence cluster Software Systems Engineering (KoSSE)
- Innovationsstiftung Schleswig-Holstein
- Excellence cluster “Future Ocean”
  - Principal Investigator, Research Area R10 “Ocean Observations”
- Currently: Dean, Faculty of Engineering

KoSSE KOMPETENZVERBUND SOFTWARE SYSTEMS ENGINEERING SCHLESWIG-HOLSTEIN

ISH-Transfer 2011 ISH-Transferprämie 2012
Prämie für engagierte Kooperation mit Unternehmen und aktiven Wissenstransfer

future ocean
KIEL MARINE SCIENCES

A brief look at Software Engineering Terminology

- Software:
  - "Computer programs, procedures, rules, and possibly associated documentation and data pertaining to the operation of a computer system."
  - IEEE Standard Glossary of Software Engineering Terminology, IEEE Standard 610.12-1990
- Concerning the term "model":
  - Software as model for the real world
  - Model of the Software
- To distinguish:
  - Software as Code
  - Software as a Service

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 …

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://www.dfg.de/en/research_funding/legal_conditions/good_scientific_practice/index.html)

"If I have seen further it is by standing on the shoulders of giants."
Isaac Newton, 1676

Not only in Germany …

EXECUTIVE OF
OFFICE OF

OPEN ACCESS RESEARCH DATA WITHOUT BARRIERS

Research Data Alliance Launch and First Plenary
March 18-20, 2013, Gothenburg, Sweden

"Sharing and cooperation are essential to science – no wonder scientists have long sought out tools to help them do this better. Remember it was scientists at CERN who invented the World Wide Web. That was a great gift of science to society: now we can ensure that it helps the scientists back."

Neelie Kroes, Vice President of the European Commission responsible for the Digital Agenda[1]

…RTMENTS AND AGENCIES

…g Access to the Results of Federally Funded Scientific Research

…olicy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

https://www.eff.org/sites/default/files/ostp-public-access-memo.pdf

Scientific misconduct and other challenges

- Prof. Dullo presented several examples of scientific misconduct, such as the case Jan Hendrik Schön.
  - Thus, I can 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 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.

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

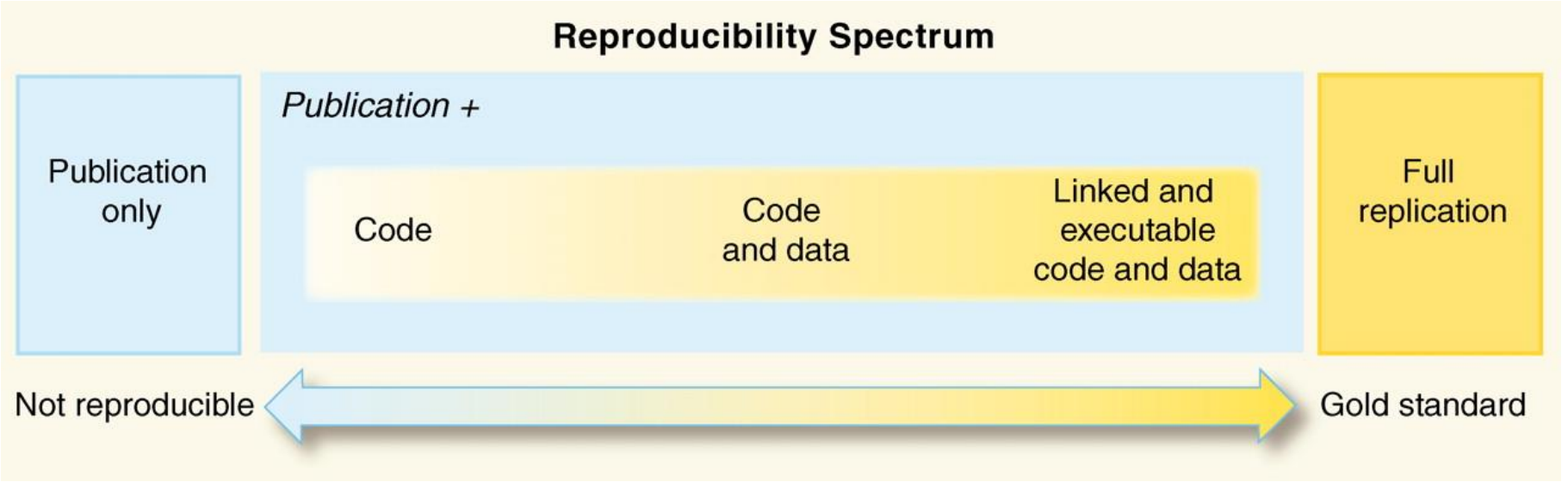
Science

PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

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



SCIENTIFIC PROGRAMMING

Editors: Konrad Hinsen, hinsen@cnrs-orleans.fr
Konstantin Läufer, laufer@cs.luc.edu

CARING FOR YOUR DATA

By Konrad Hinsen

http://dx.doi.org/10.1109/MCSE.2012.108

"Data is at the heart of science.

- A scientist is **expected** to be able to back up all published conclusions with data.
- **Data management** should thus be a priority in science.
- Scientists can't afford to lose data, be uncertain of what it means, or not know where it came from.
- In the **experimental sciences**, there's a long tradition of writing down all experimental setups, parameters, and results meticulously in a lab notebook.
- Unfortunately, **computational science** is much less rigorous about data handling, although there are clear signs of improvement."

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:// www.plagiarismfinder.de/

http://vroniplag.wikia.com/

VroniPlag

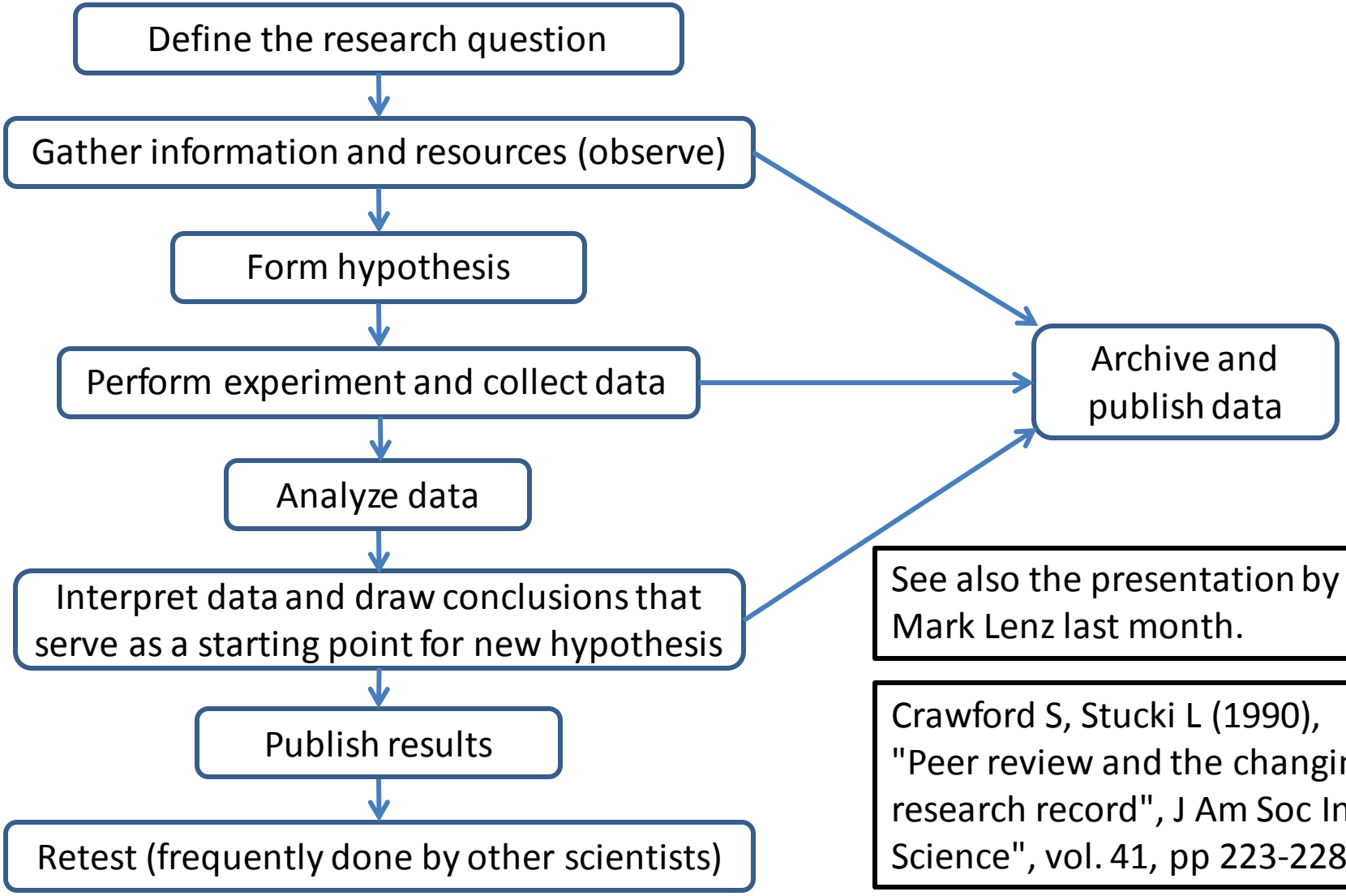
What I'm not talking about?

- Establishing Software Engineering best practices in computational science, such as
  - Version and configuration management
  - Quality management
  - Software architecture design and modeling
  - Domain-specific programming languages
  - 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



See also the presentation by Mark Lenz last month.

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

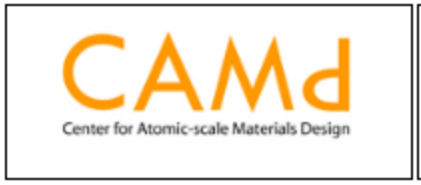
World Data Center for Marine Environmental Sciences

Biogeochemistry, Circulation, and Life of Present and Past Oceans

http://www.wdc-mare.org/

MODEL & DATA

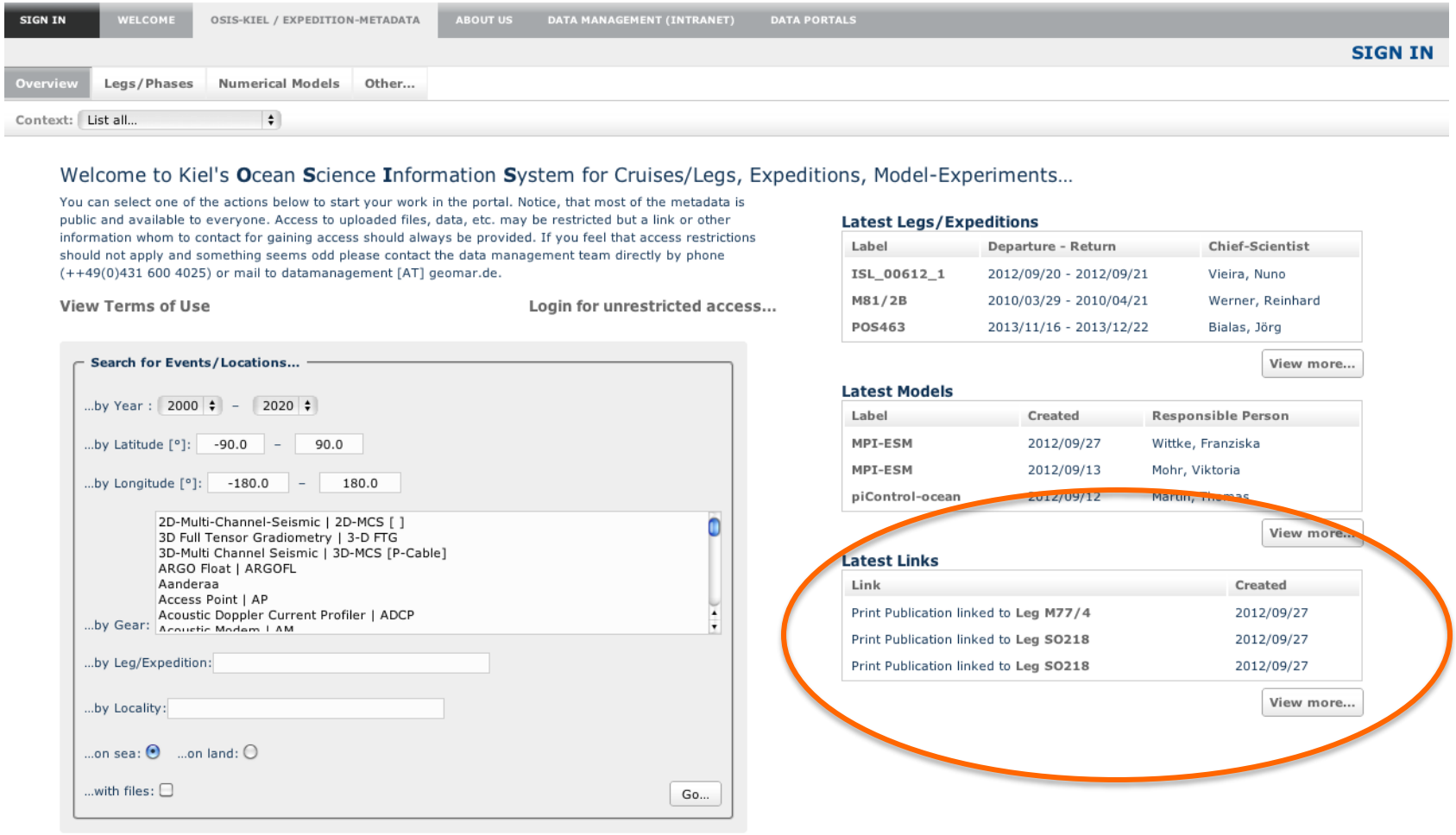
http://www.mad.zmaw.de/wdc-for-climate/



COMPUTATIONAL MATERIALS REPOSITORY

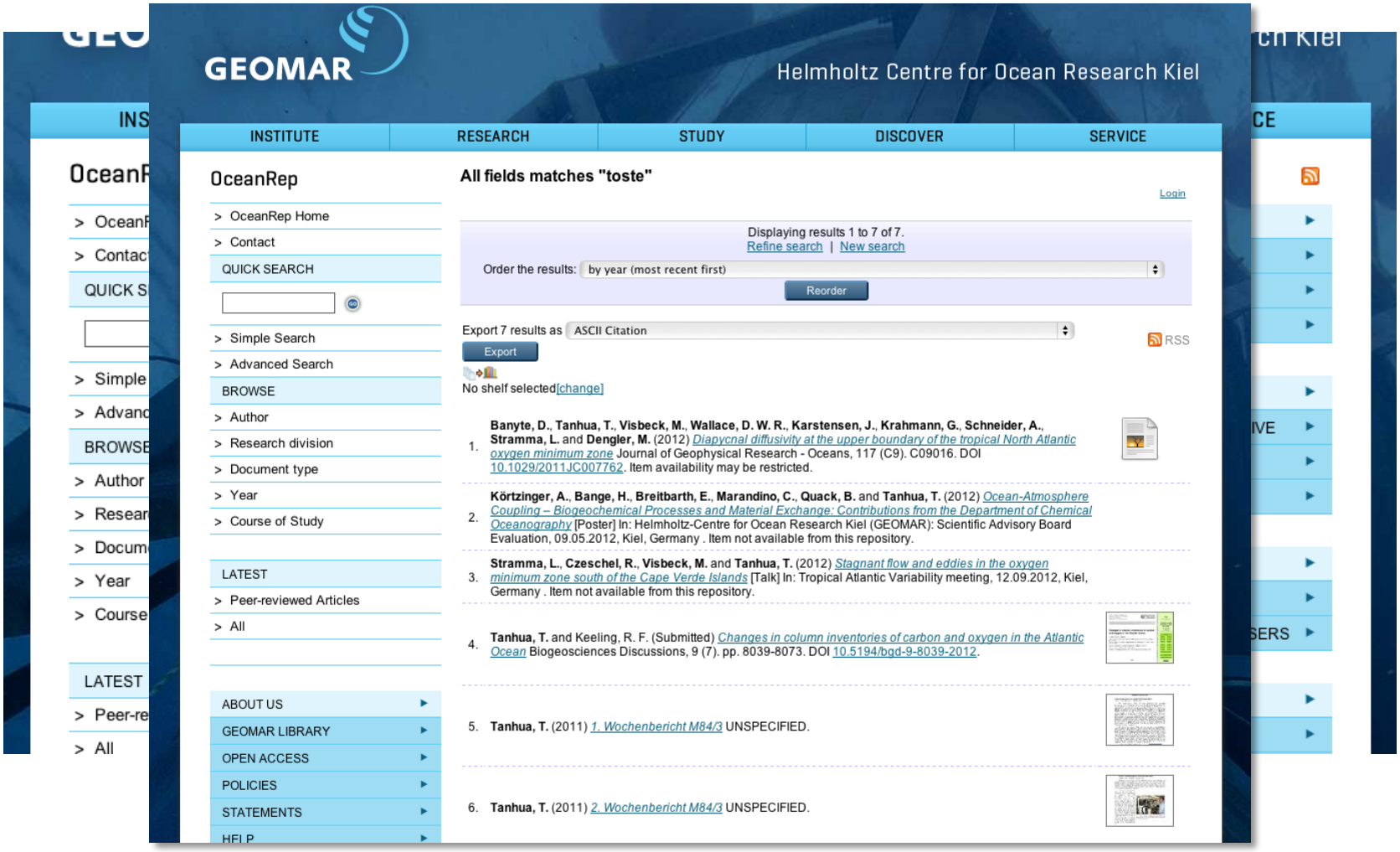
https://cmr.fysik.dtu.dk/

Kiel Data Management Infrastructure – Ocean Science Information System

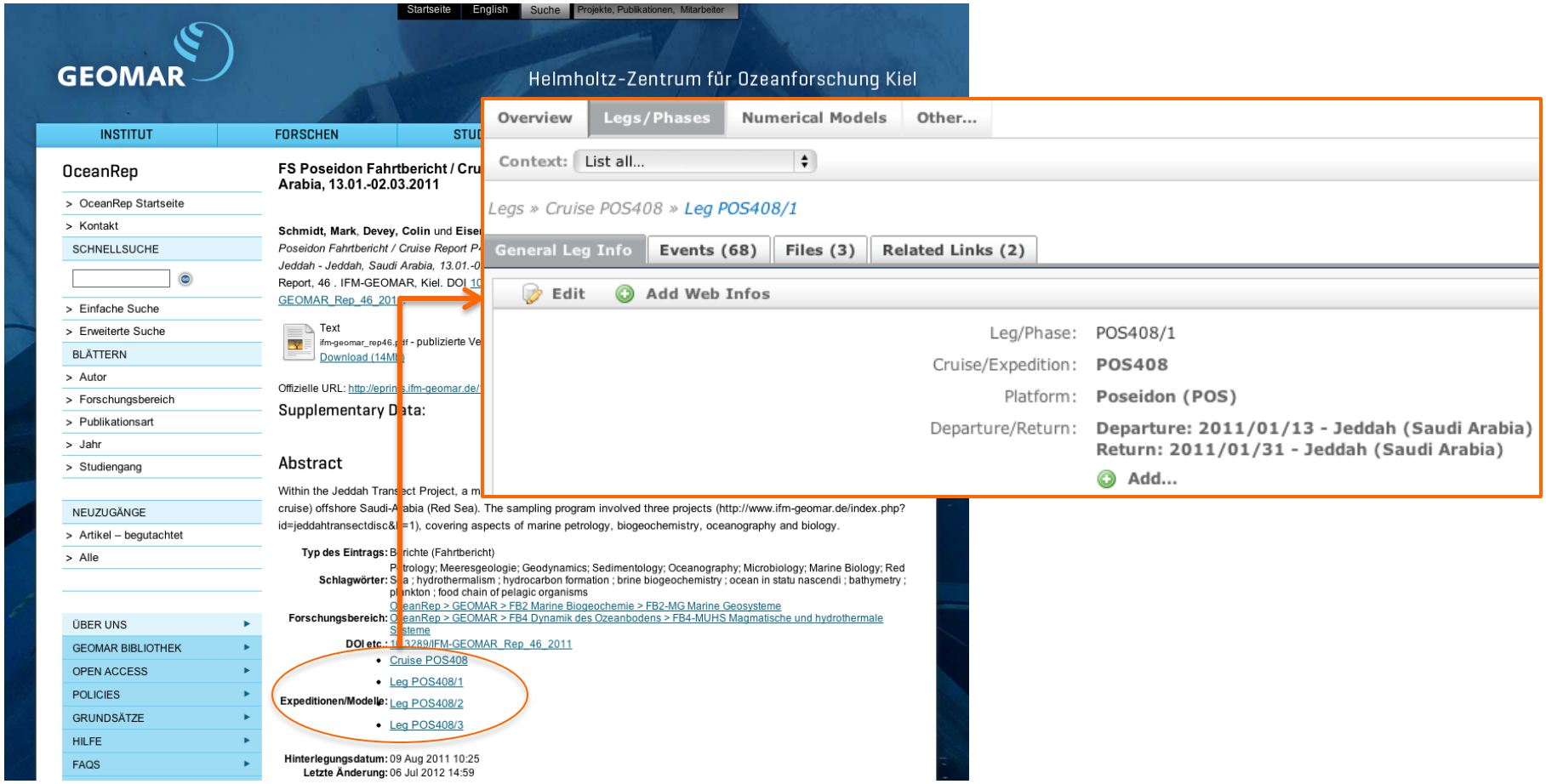


Source: Kiel Data Management Team, https://portal.geomar.de/web/guest/kdmi

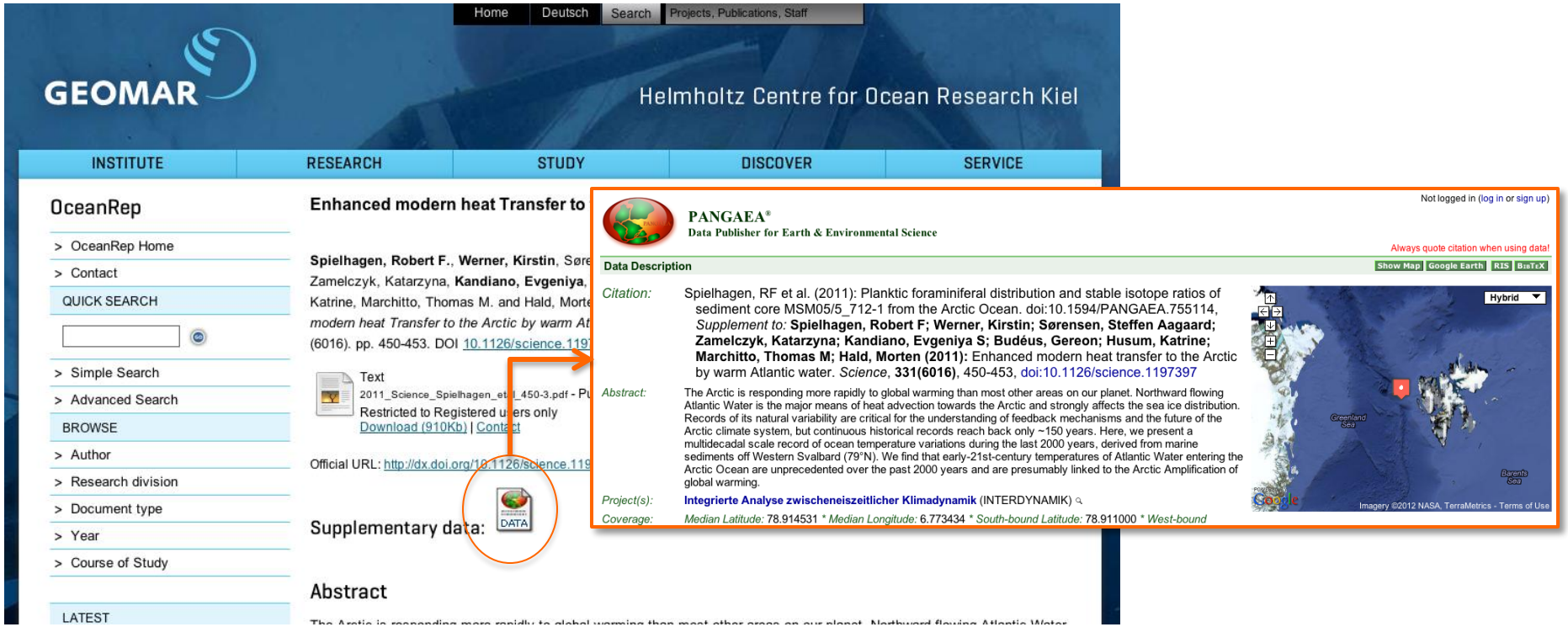
GEOMAR – OceanRep



GEOMAR – OceanRep link to data



GEOMAR – OceanRep link to Pangaea



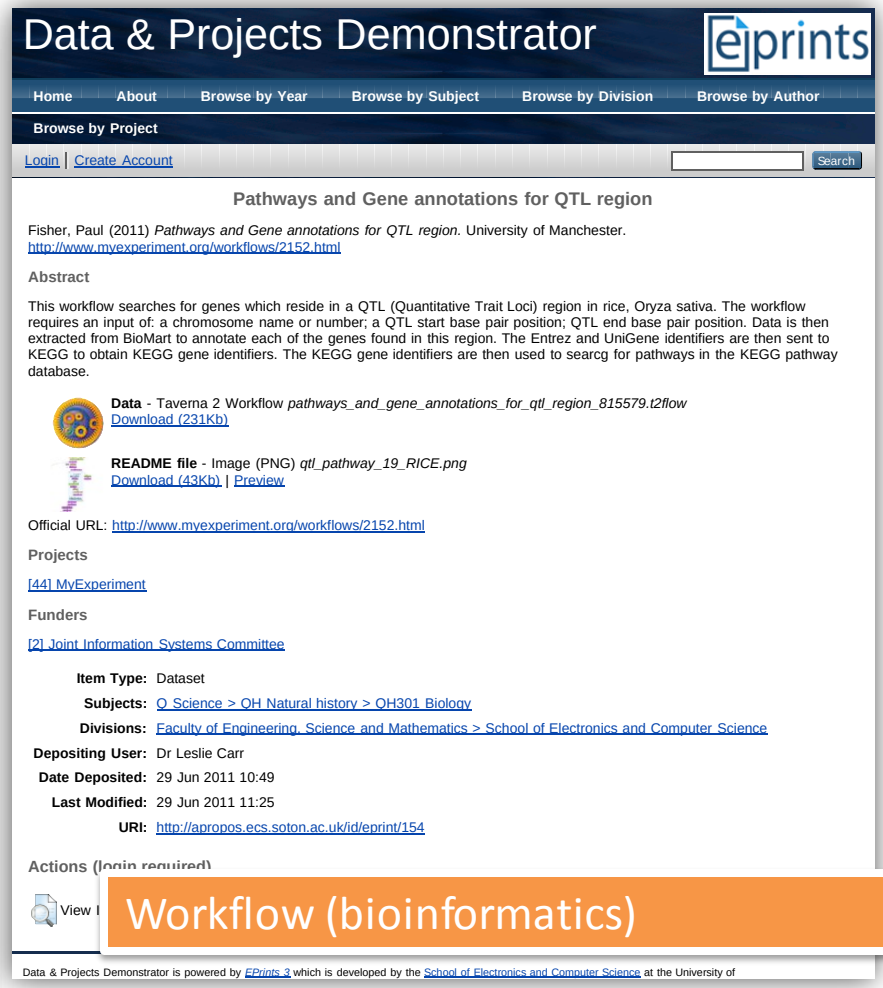
Kielprints is a similar service for Kiel at large:

http://eprints.uni-kiel.de

Recently, a collaboration with the Eprints Group at the University of Southampton started.

EPrints 3.3 for Research Data

Experiment (crystallography)



Workflow (bioinformatics)

Source: Lesly Carr, University of Southampton

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:

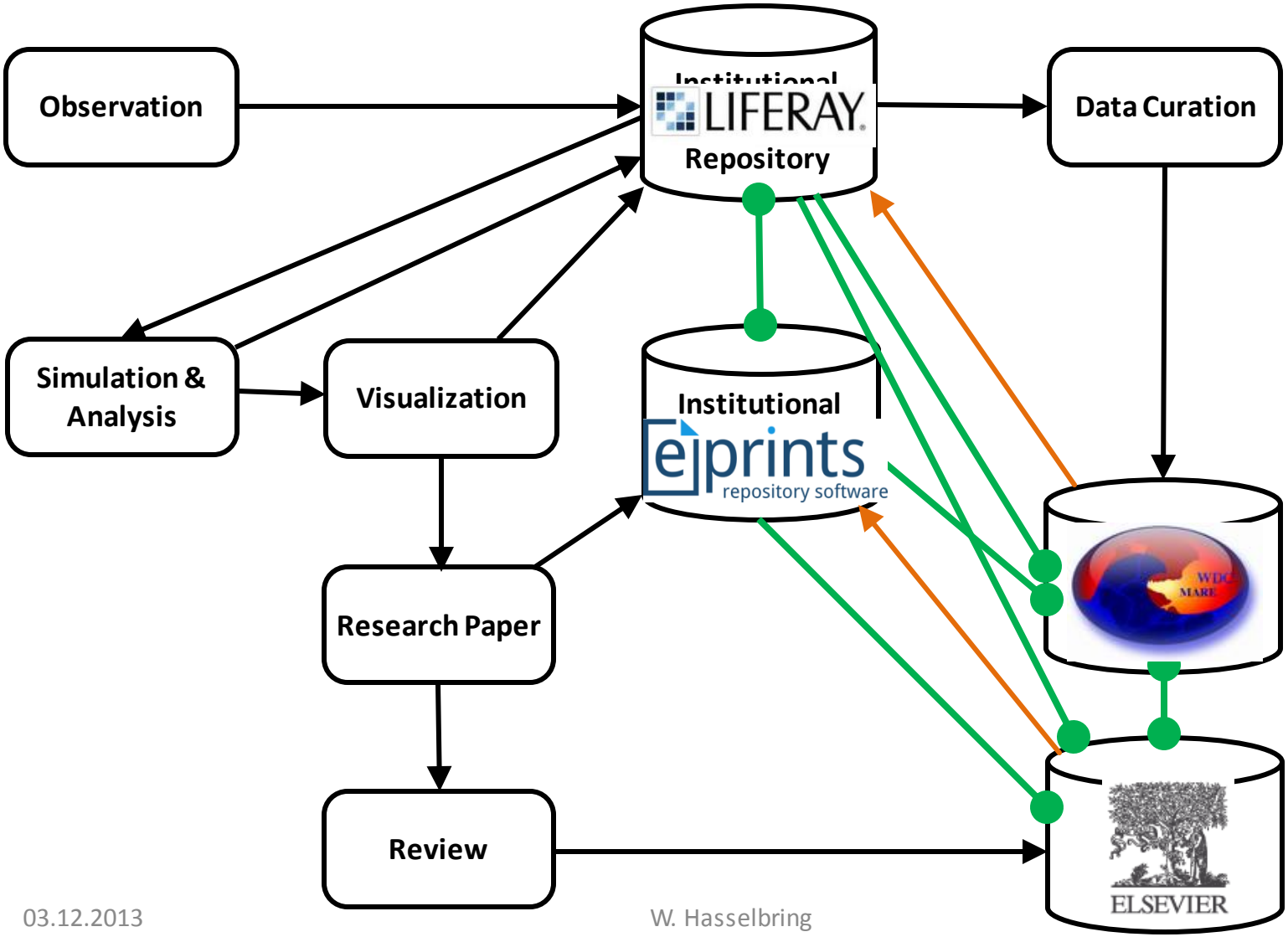
Software Engineering Group, University Kiel

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/

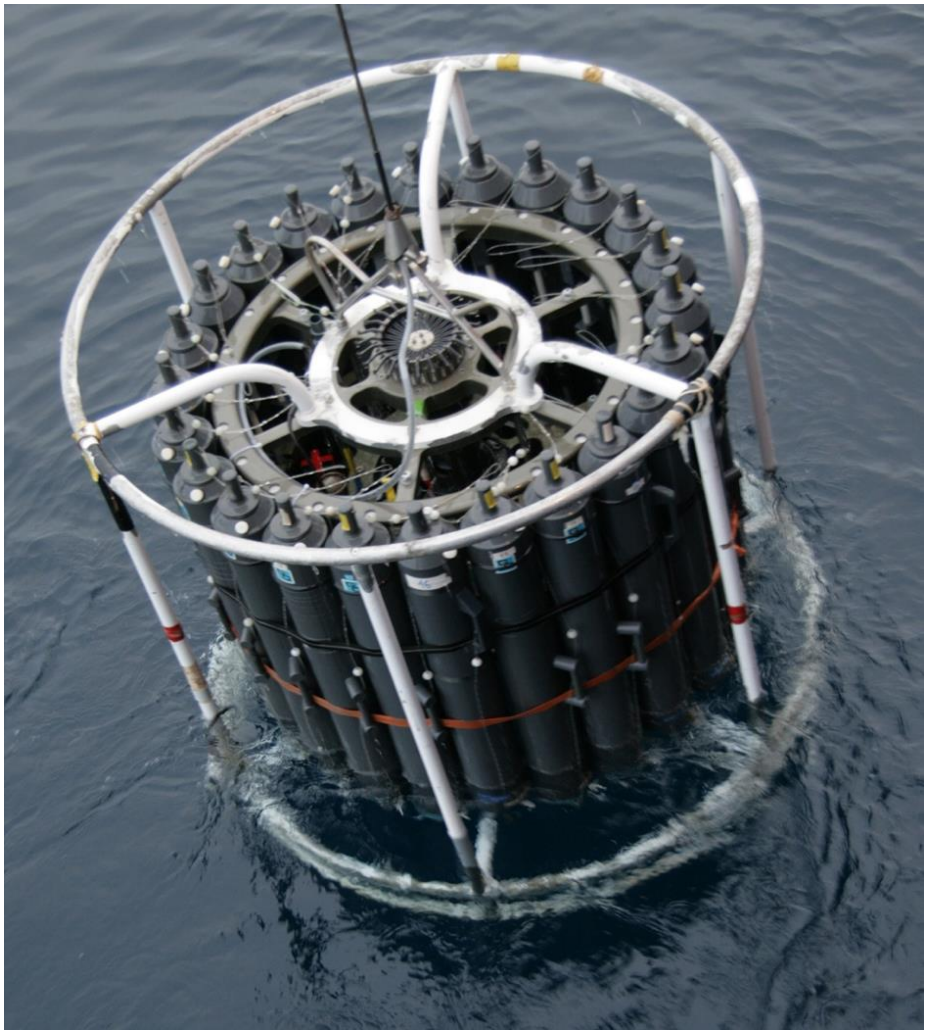
Data and Paper Flow (in Ocean Science)



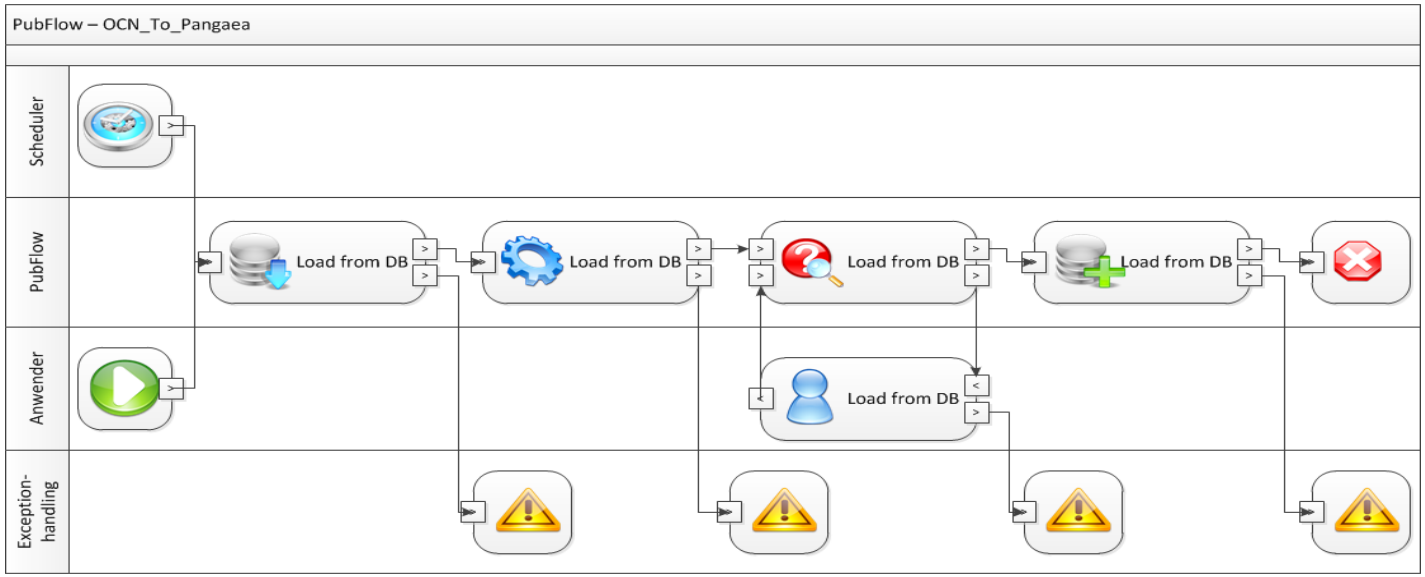
The initial PubFlow case study

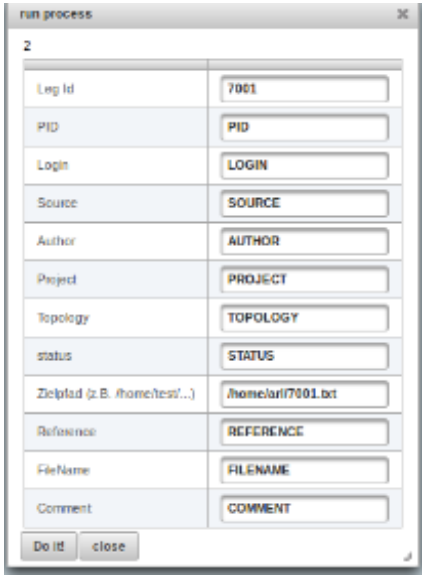
CTD Rosette System

(Conductivity / Temperature / Depth/pressure sensor).



CTD Workflow





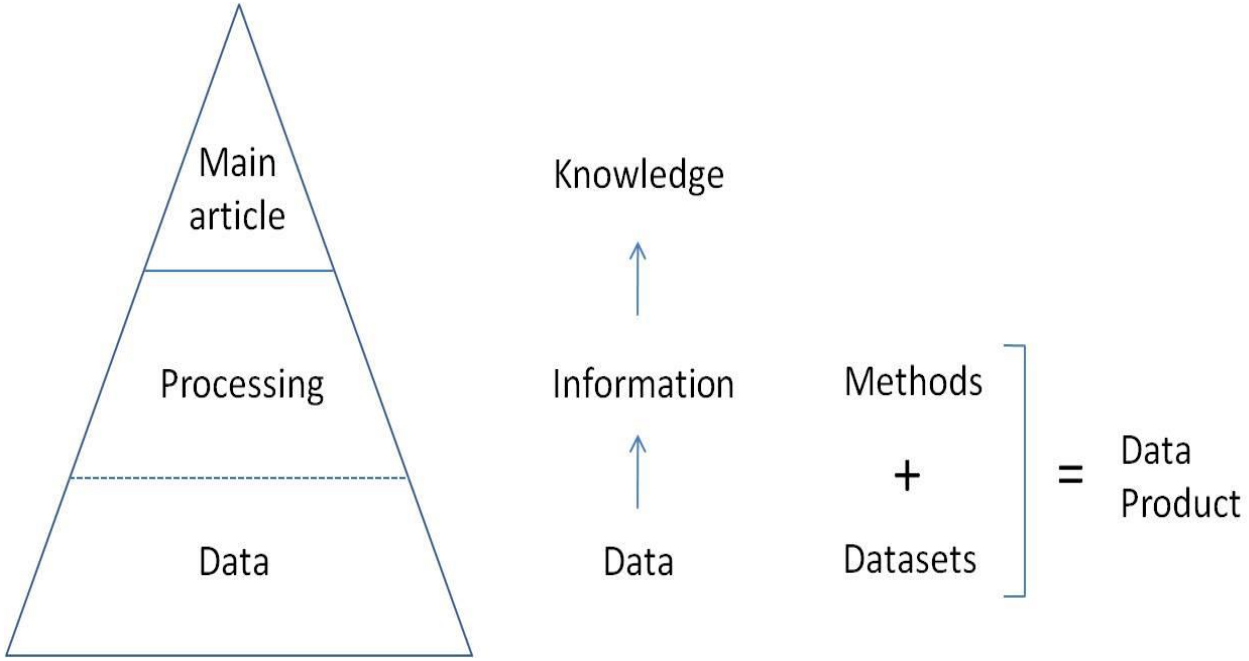
JBoss® jBPM

JIRA

Automatically collecting provenance data in PubFlow

- Provenance information describes the origins and the history of (research) data in its life cycle, and the process by which it arrived:
  - A description of all modifications and changes
  - A list of all methods applied
  - A list of all systems and persons involved
  - Answers to
    - when,
    - where,
    - how,
    - why.
- Given the documented history of an object, the object attains an **authority** that allows scholars to appreciate its importance, whereas, in the absence of such history, the object may be treated with some skepticism.
- Usually collected and used in databases, data warehouses, workflow systems, scientific data processing, Web services, etc.
- We employ Kieker [Rohr et al. 2008, van Hoorn et al. 2009] for automatically collecting provenance data [Brauer and Hasselbring 2012, 2013a, 2013b].

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

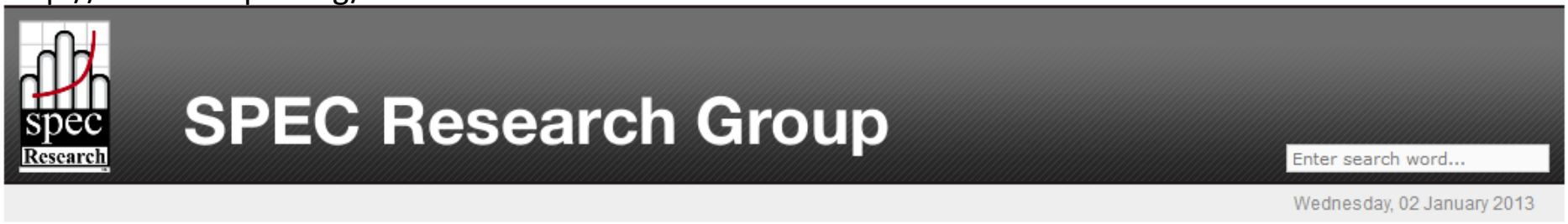
doi:10.1038/nature10836

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."

http://research.spec.org/



- Home
- News
- FAQ
- Mission and Charter
- Working Groups
- ICPE Proceedings
- Benchmarking Portal
- Awards
- **Tools**
  - Submission
- Newsletter
- Member Organizations
- Steering Committee
- Contact

Quantitative Evaluation and Analysis Tools

This page hosts a repository of peer-reviewed tools for quantitative system evaluation and analysis. The published tools have undergone a thorough review process by multiple independent experts to ensure high quality and relevance to the community. The review process covers important quality factors, including maturity, availability and usability.

Overview of Hosted Tools

Tool name	Description	License	Requirements
Faban	Faban is a framework and tool to automate the creation and running of multi-tier server performance tests/workloads.	CDDL 1.0	Unix or Windows. Java 5 or later.
Kieker	Kieker is a framework for application performance monitoring and dynamic software analysis.	Apache License, Version 2.0	Java 5 or later.

Important Links

- RG Cloud Working Group
- IDS Benchmarking Working Group
- SPEC Distinguished Dissertation Award
- Quantitative Evaluation and Analysis Tools
- Latest Newsletter Issue
- ICPE International Conference

See, for instance: [Waller and Hasselbring 2012]

Source code and **benchmark** code / data are available at http://kieker-monitoring.net/

Executable, data papers

More info at
http://www.executablepapers.com/
http://www.articleofthefuture.com/

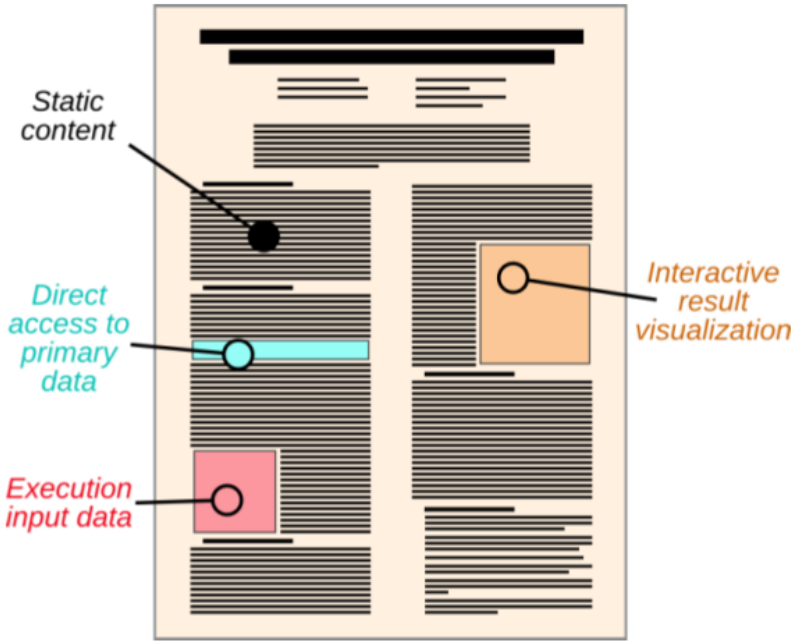
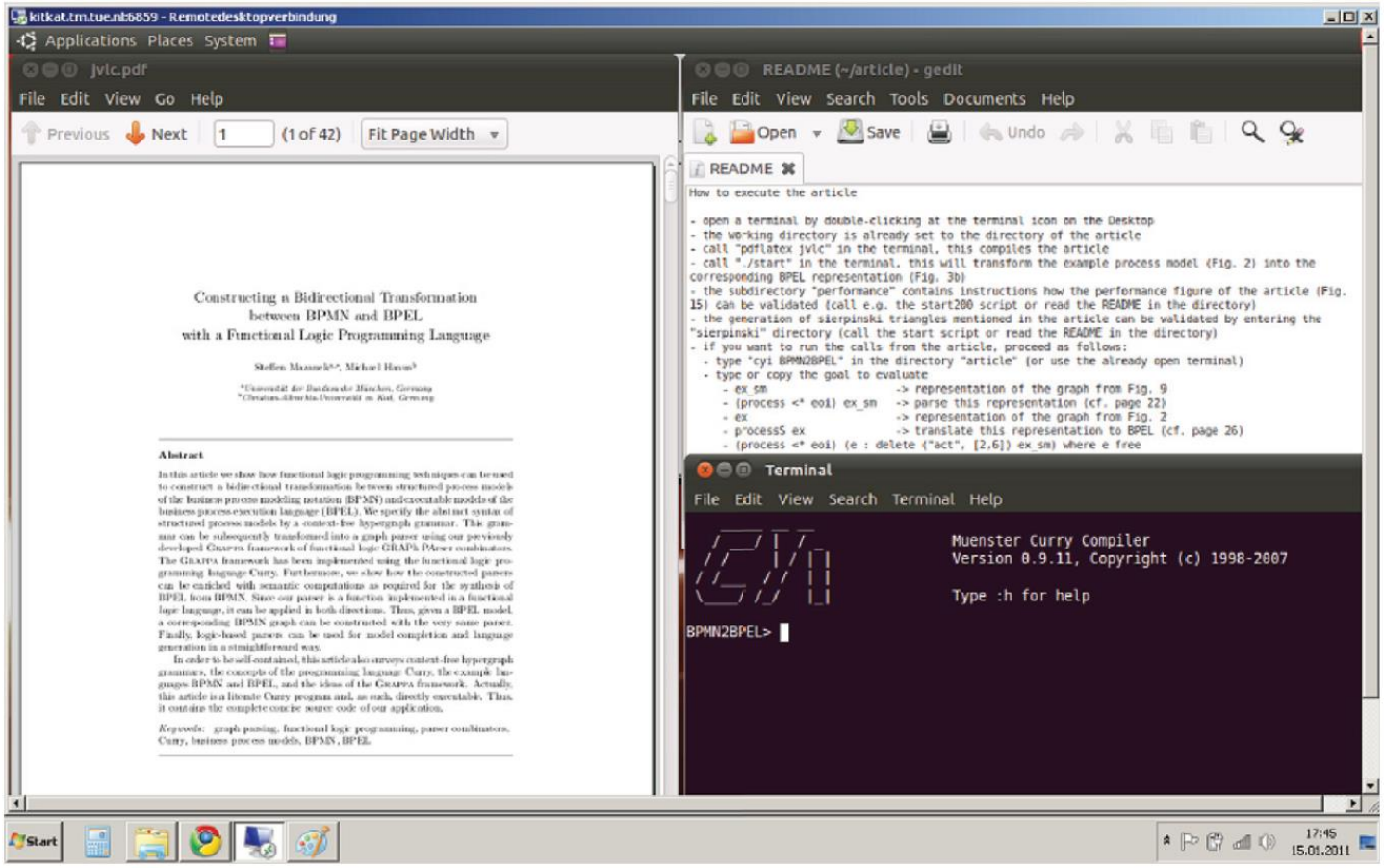


Fig. 1: Conceptual view of the executable paper. Static content (the body of the publication) is extended by interactiv
access primary data and reenact computations in order to validate the presented conclusions or navigate result spaces
approval, readers can also obtain access to the underlying code of the experiments presented in the publication. The i
and can be integrated with the Publisher's portal.

An Example Executable Paper



SHARE: a web portal for creating and sharing executable research papers
http://sites.google.com/site/executablepaper/
http://dx.doi.org/10.1016/j.procs.2011.04.062
http://www.springerlink.com/content/y148817864016412/

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

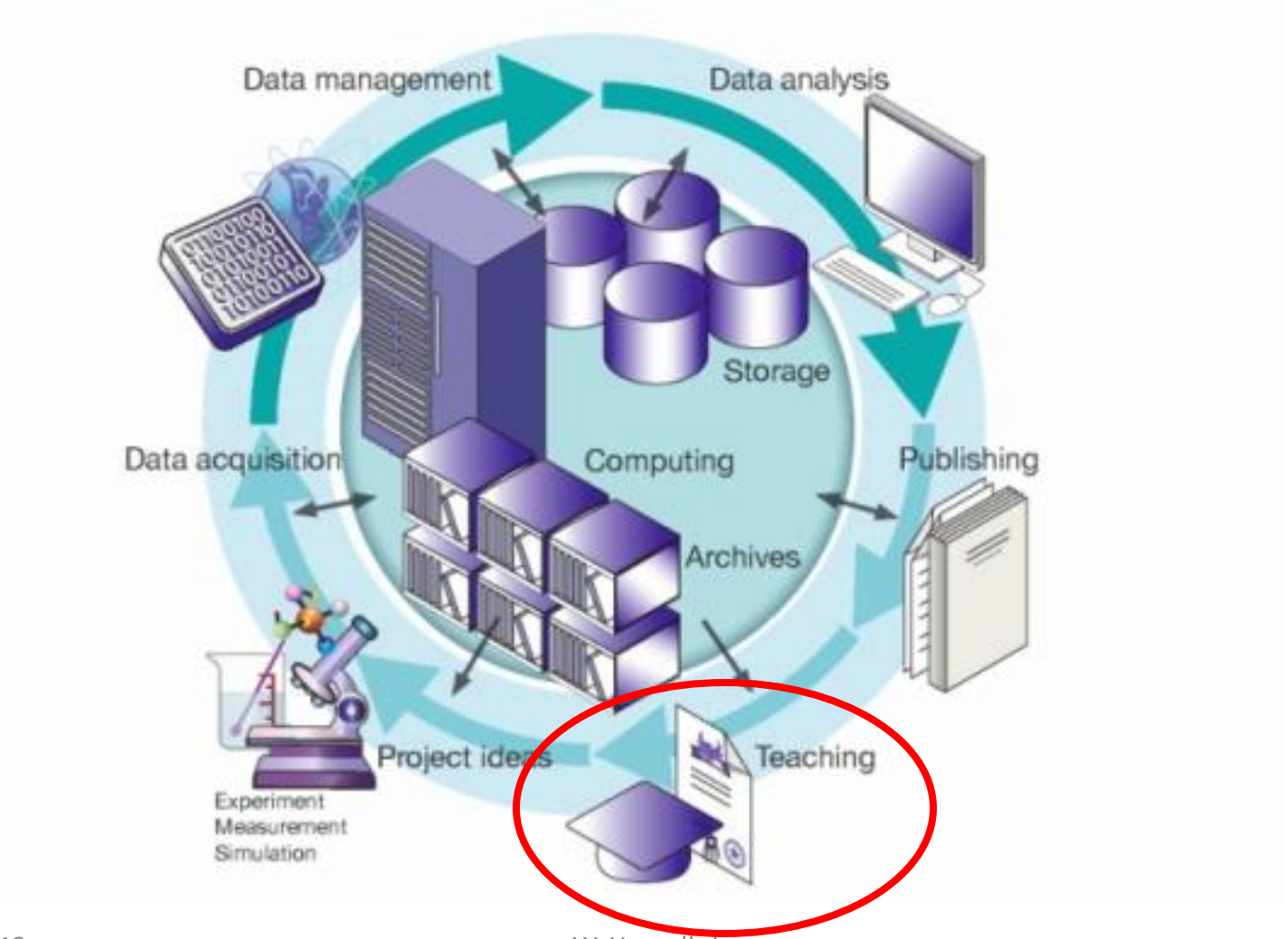
Metrics scenarios for the data publishing models
(and comparison with the current scientific publication model)

Types of metrics	Currently available tools with possibilities for "data metrics"	Metric dimensions	Models			
					Data Publication	
			Scientific publication	Data publication	Stand-alone data publications	Journal data publications
Data publication & citation-based metrics	- Data Citation Index (Web of Science) - Google Scholar - Scopus - Microsoft Academic Search - DataCite	Size-dependent	Yes	Difficult (1)	Yes (4)	Yes
		Size-independent				
		- Direct average performance	Yes	No	Yes	Yes
		- Source-based performance	Yes	No	Yes (3)	Yes
Altmetrics-based metrics	- ImpactStory - Twitter, Facebook	Social media indicators	Yes	Yes	Yes	Yes
	- Mendeley - CiteULike	Readership counts	Yes	No	Yes	Yes
	- Repositories - Data Journals	Downloads & views counts (DUI metrics)	Yes	Difficult (2)	Yes	Yes

http://www.knowledge-exchange.info/datametrics

The data life cycle

(http://www.scc.kit.edu/forschung/lsdf.php)



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

References

[Brauer and Hasselbring 2012] Brauer, P. C. and Hasselbring, W. “Capturing provenance information with a workflow monitoring extension for the Kieker framework.” In: The Third International Workshop on the role of Semantic Web in Provenance Management , 28.5.2012, Heraklion, Kreta. http://eprints.uni-kiel.de/19636/

[Brauer and Hasselbring 2013a] Brauer, P. C. and Hasselbring, W. “PubFlow: a scientific data publication framework for marine science.” In: IMDIS 2013 International Conference on Marine Data and Information Systems, 23.-25.09.2013, Lucca, Italy. http://eprints.uni-kiel.de/22399/

[Brauer and Hasselbring 2013b] Brauer, P. C. and Hasselbring, W. (2013) “PubFlow: provenance-aware workflows for research data publication.” In: 5th USENIX Workshop on the Theory and Practice of Provenance (TaPP '13), April 2-3, 2013, Lombard. http://eprints.uni-kiel.de/21112/

[Rohr et al. 2008] M. Rohr, A. van Hoorn, J. Matevska, N. Sommer, L. Stoever, S. Giesecke, and W. Hasselbring. “Kieker: Continuous monitoring and on demand visualization of Java software behavior.” In C. Pahl, editor, Proceedings of the IASTED International Conference on Software Engineering 2008 (SE’08), pages 80–85, Feb. 2008. ISBN 978-0-88986-715-4. http://eprints.uni-kiel.de/14496/

[van Hoorn et al. 2009] A. van Hoorn, M. Rohr, W. Hasselbring, J. Waller, J. Ehlers, S. Frey, and D. Kieselhorst. Continuous monitoring of software services: Design and application of the Kieker framework.” Technical Report TR-0921, Department of Computer Science, University of Kiel, Germany, Nov. 2009b. http://eprints.uni-kiel.de/14459/

[van Hoorn et al. 2012] A. van Hoorn, J. Waller, and W. Hasselbring. Kieker: A framework for application performance monitoring and dynamic software analysis. In Proceedings of the 3rd ACM/SPEC International Conference on Performance Engineering (ICPE 2012), pages 247–248. ACM, Apr. 2012. ISBN 978-1-4503-1202-8. http://eprints.uni-kiel.de/14418/

[Waller and Hasselbring 2012] J. Waller and W. Hasselbring: “A Comparison of the Influence of Different Multi-core Processors on the Runtime Overhead for Application-Level Monitoring.” Springer LNCS vol. 7303, 2012, http://eprints.uni-kiel.de/15425/