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
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
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.”

• 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 ...
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 ...

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

Policy 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

03.12.2013
W. Hasselbring
6
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

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

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 and the input/configuration data of the model, reproducibility of the results is hard or even impossible.
“Replication is the ultimate standard by which scientific claims are judged.”
“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.”

http://dx.doi.org/10.1109/MCSE.2012.108
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://www.plagiarismfinder.de/

http://vroniplag.wikia.com/
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

Define the research question

Gather information and resources (observe)

Form hypothesis

Perform experiment and collect data

Analyze data

Interpret data and draw conclusions that serve as a starting point for new hypothesis

Publish results

Retest (frequently done by other scientists)

Archive and publish data

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/

Modelle & Daten

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

Center for Atomic-scale Materials Design

https://cmr.fysik.dtu.dk/
Welcome to Kiel’s Ocean Science Information System for Cruises/Legs, Expeditions, Model-Experiments...

You can select one of the actions below to start your work in the portal. Notice, that most of the metadata is public and available to everyone. Access to uploaded files, data, etc. may be restricted but a link or other information whom to contact for gaining access should always be provided. If you feel that access restrictions should not apply and something seems odd please contact the data management team directly by phone (+49(0)431 500-4025) or mail to data-management [AT] geomar.de.

View Terms of Use
Login for unrestricted access...

Latest Legs/Expeditions

<table>
<thead>
<tr>
<th>Label</th>
<th>Departure - Return</th>
<th>Chief-Scientist</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISL_00612_1</td>
<td>2012/09/20 - 2012/09/21</td>
<td>Viela, Nuno</td>
</tr>
<tr>
<td>MBI/2B</td>
<td>2010/09/29 - 2010/04/21</td>
<td>Werner, Reinhard</td>
</tr>
<tr>
<td>POS463</td>
<td>2013/11/16 - 2013/12/22</td>
<td>Bielas, Jorg</td>
</tr>
</tbody>
</table>

Latest Models

<table>
<thead>
<tr>
<th>Label</th>
<th>Created</th>
<th>Responsible Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPI-ESM</td>
<td>2012/09/27</td>
<td>Wittke, Franziska</td>
</tr>
<tr>
<td>MPI-ESM</td>
<td>2012/09/13</td>
<td>Mohr, Viktoria</td>
</tr>
<tr>
<td>pControl-ocean</td>
<td>2012/09/12</td>
<td>Haake, Thomas</td>
</tr>
</tbody>
</table>

Latest Links

<table>
<thead>
<tr>
<th>Link</th>
<th>Created</th>
</tr>
</thead>
<tbody>
<tr>
<td>Print Publication linked to Leg M77/4</td>
<td>2012/09/27</td>
</tr>
<tr>
<td>Print Publication linked to Leg SO218</td>
<td>2012/09/27</td>
</tr>
<tr>
<td>Print Publication linked to Leg SO218</td>
<td>2012/09/27</td>
</tr>
</tbody>
</table>

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

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
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.

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

$\text{Algorithms + Data Structures} = \text{Programs}$

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.”
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 interactive access to primary data and reenact computations in order to validate the presented conclusions or navigate result spaces. After approval, readers can also obtain access to the underlying code of the experiments presented in the publication. The integration can be integrated with the Publisher’s portal.

03.12.2013
W. Hasselbring
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/y1488178640l6412/
What about social networks?

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

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

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)

<table>
<thead>
<tr>
<th>Types of metrics</th>
<th>Currently available tools with possibilities for “data metrics”</th>
<th>Metric dimensions</th>
<th>Scientific publication</th>
<th>Data Publication</th>
<th>Data Stand-alone publications</th>
<th>Journal data publications</th>
</tr>
</thead>
</table>
| 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  
- Mendeley  
- CiteULike  
- Repositories  
- Data Journals | Social media indicators | Yes | Yes | Yes | Yes |
| | | Readership counts | Yes | No | Yes | Yes |
| | | 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


