Research Compendia: Connecting Computation to Publication

Jennifer Seiler

Department of Statistics
Columbia University
New York, NY

Scientific Software Days
Austin, TX
December 16, 2013

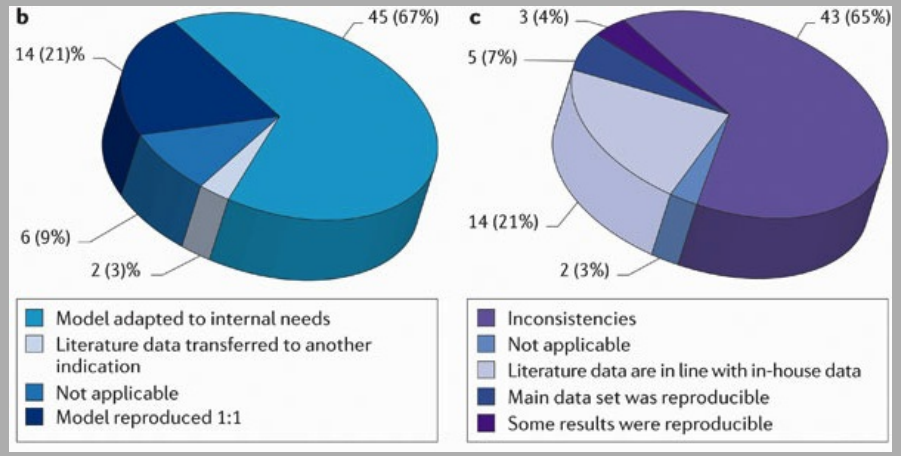
Outline

Reproducibiltity

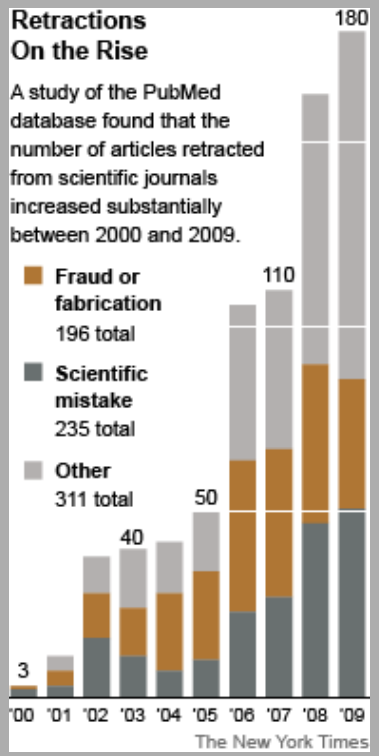
"The idea is: An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete ... set of instructions [and data] which generated the figures." David Donoho, 1998.

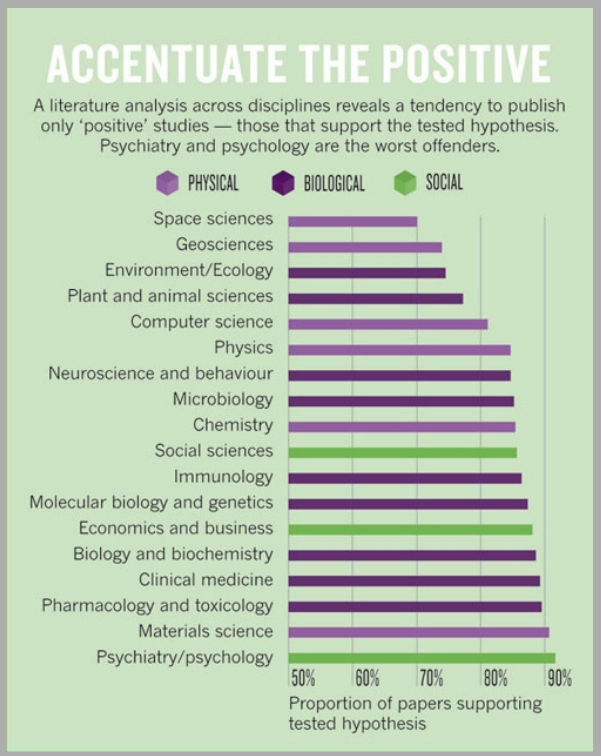
1. Reproducibility is a keystone of the scientific method.
   - It is essential for credibility, for the scientific record, and for efficient scientific progress.
2. Openness/Transparency is essential for reproducibility.

The Crisis of Credibility

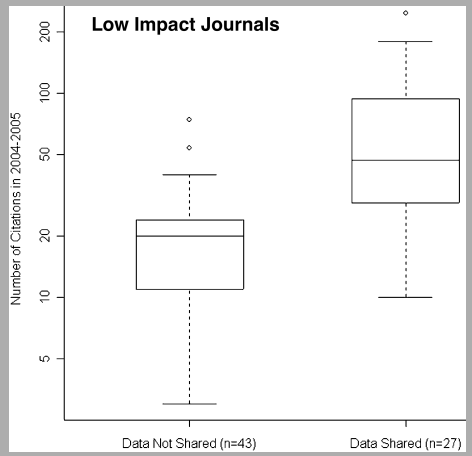


- Decline in result significance and reproducibility is rearing its head in multiple fields. (30% reproducibility in psychology [1], 21% in drug trials [2], 11% in new cancer drug studies [3]).
- Irreproducibility is the biggest factor in the ever increasing numbers of retractions [4].
- Young researchers are being trained to hide bad results [5].



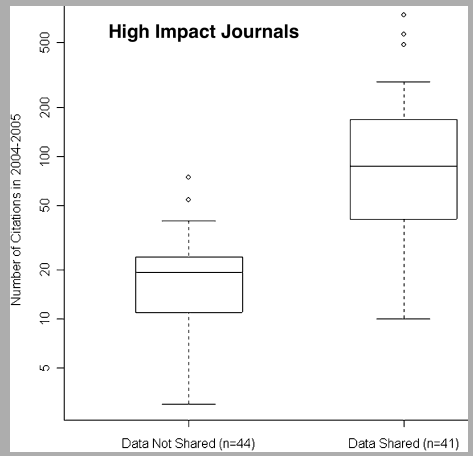


Why share?



	Number of Articles		
	Total	Data Shared	Data Not Shared
TOTAL	85	41 (48%)	44 (52%)
High Impact (>=25)	12	12 (100%)	0 (0%)
Low Impact Journal	73	29 (40%)	44 (60%)
Published 1999–2000	6	5 (83%)	1 (17%)
Published 2001–2003	79	36 (46%)	43 (54%)
Include a US Author	56	35 (63%)	21 (38%)
No US Authors	29	6 (21%)	23 (79%)

doi:10.1371/journal.pone.0000308.t001



Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). “Sharing detailed research data is associated with increased citation rate.” PloS one, 2(3), e308.

- Boost your citation rates and recognition, and get credit for your code in addition to your publications.
- Defend your research.
- Archive your research and track your procedures.
- Speed up the process of converting scientific results into productive forces.
- Allow others to build on your work in a highly visible and trackable way.
- Journals and funding agencies are implementing policies requiring sharing of data and code.

The Cultural War

If it benefits everyone, why doesn't everyone share?

- Time and resource pressures:
  - 'We don't have time to organize and collect our data and code in a readable way'
  - 'The code is messy'/'The data isn't labeled in a readable way for the public'
- Copyright fears and the fear of being scooped:
  - 'We are still using it'/ 'It is only for collaborators'
  - 'We cannot send our code out of house'
- Social/Institutional:
  - 'We would have to get consensus from all of our collaborators to share that code'
- 'Meincraft'
  - 'I worked hard on that code/data and do not wish to make it public'
- Before there was staggering evidence to show that sharing increased visibility and citations for research, scientists have historically been offered little to no incentive to share.

Reasons for not sharing:

doi:10.1001/jama.287.4.473

What is "Open"?

> *"All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader. All computer codes involved in the creation or analysis of data must also be available to any reader of Science. After publication, all reasonable requests for data and materials must be fulfilled."* ***AAAS Science Magazine submission policy***

- Every field is different
- There are many legitimate obstacles to sharing code and data:
  - Size constraints (i.e. many terabytes)
  - Sensitive data (e.g. human studies)
  - Special circumstances (e.g. designed to run on custom proprietary firmware)
- The above describes *reviewable research*. That is, it can be independently assessed and the results judged credible, but it does not necessarily imply reproducibility.
- More is needed to provide for reproducibility, i.e. metadata.

Best Practices

To provide for reproducibility we need to have [7]:

- The extent of computational work to be performed.
- Platforms and software to be utilized.
- Thorough dataset and software documentation.
- Reasonable standards for persistence of resulting software and dataset preservation and archiving.
- Full disclosure of salient details regarding software and data use including:
  - specification of the dataset used
  - details of the algorithms employed
  - the hardware and software environment, versions, and environmental variables
  - parameters and scripts used
  - the testing performed/commands run, workflow tracking, etc.
- There is also a need for better standards on how to include citations for software and data in the references of a paper. Not inline or as footnotes.

The Status Quo

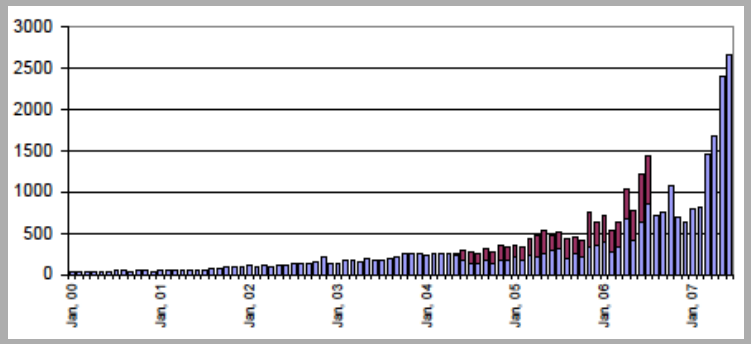
Best practices training is, for the most part, nonexistent in most of academic research:

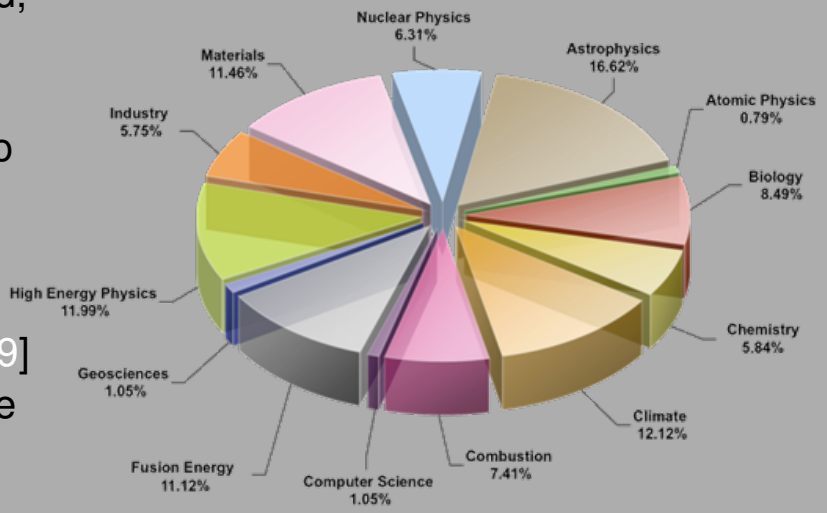
- Scientists are currently not taught how to produce reproducible research.
- Further, many scientists never learn good computational science techniques, such as consistency and convergence tests, numerical precision, listing variable arbitrary coding choices . . .
- Code and data are still usually only referenced within supplemental texts or footnotes, if at all.
- Lack of description of computational hardware, environmental variables, compiler versions, or even common software used (e.g. Excel or Matlab)
- Parameter files and scripts are considered disposable. Even though analogous tabletop experimental steps would be considered important to preserve in lab notebooks.

The Status Quo

Meanwhile computation is becoming central to scientific research . . .

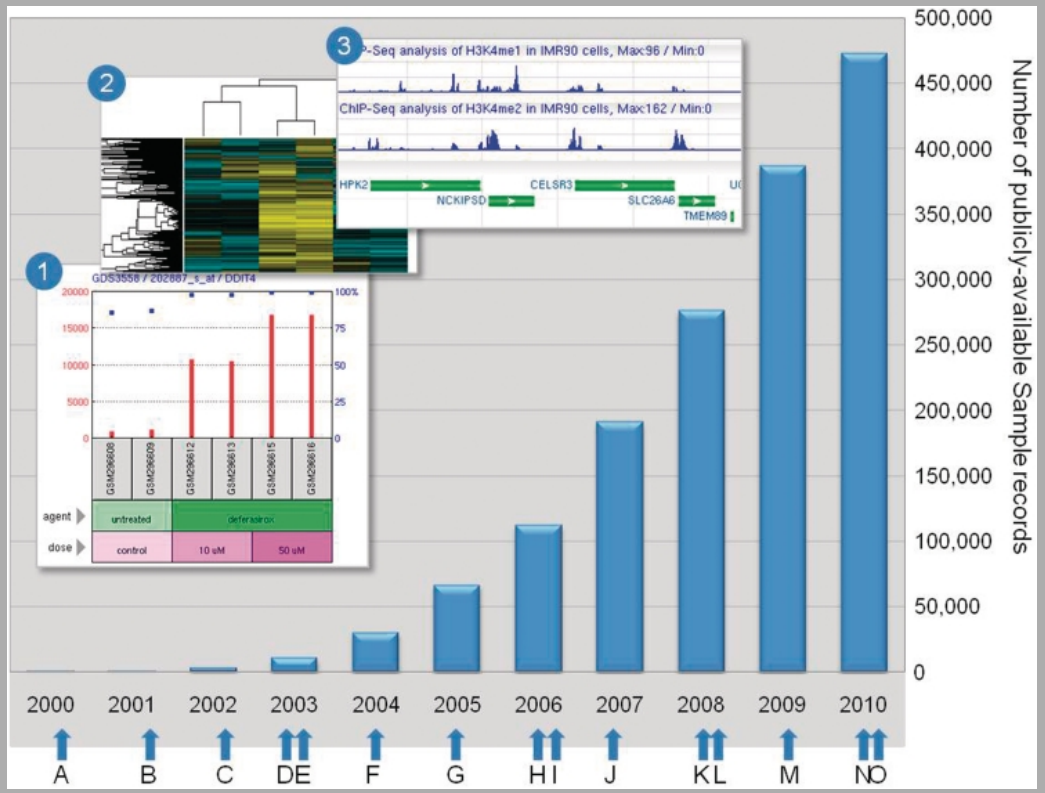
- there are enormous, and increasing, amounts of data collection:
  - CMS project at LHC: 150MB per second = 780TB/yr => several PB when data processed,
  - Sloan Digital Sky Survey: 9th data release (SDSS-III 2012), 60TB,
  - quantitative revolution in social science due to abundance of social network data [8]
  - Science survey of peer reviewers: 340 researchers regularly work with datasets >100GB; 119 regularly work with sets >1TB [9]
  - 90% of world's research data generated in the last 2 year [10]

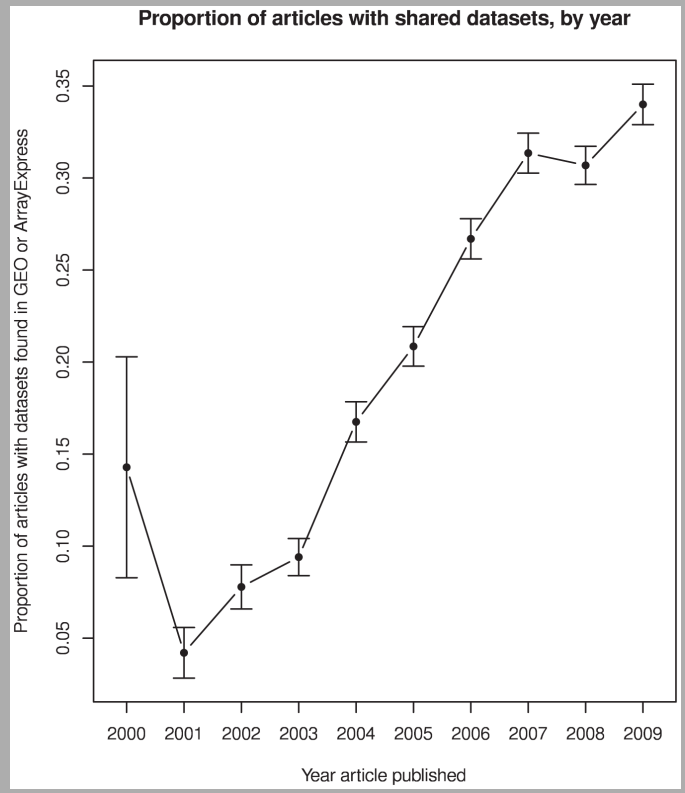




- large datasets covering higher dimensional parameter spaces require more advanced (computational) analysis,
- increasingly massive simulations of mathematical, physical and biological systems are utilized in all fields
- and deep theoretical contributions are increasingly now encoded in software.

Data Sharing





As more and more journals, governments, and institutions are adopting data sharing policies, the culture is changing and more people are volunteering their data, independent of policy.

Code Sharing

JASA June	Computational Articles	Code Publicly Available
1996	9 of 20	0
2006	33 of 35	9
2009	32 of 32	16
2011	29 of 29	21

Generally, code not available at the time of publication, insufficient information in the publication for verification of results. As you can see, far behind data.

Lists of online repositories: re3data.org (623), databib.org (602)
. . . 17 of those support code sharing

Journal Policy

Data Sharing Policy	2011	2012	Change
Required as condition of publication, barring exceptions	18	19	1
Required but may not affect editorial decisions	3	10	7
Encouraged/addressed, may be reviewed and/or hosted	35	30	-5
Implied	0	5	5
No mention	114	106	-8

Code Sharing Policy	2011	2012	Change
Required as condition of publication, barring exceptions	6	6	0
Required but may not affect editorial decisions	6	6	0
Encouraged/addressed, may be reviewed and/or hosted	17	21	4
Implied	0	3	3
No mention	141	134	-7

Source: Stodden, Guo, Ma (2013) PLoS ONE, 8(6)

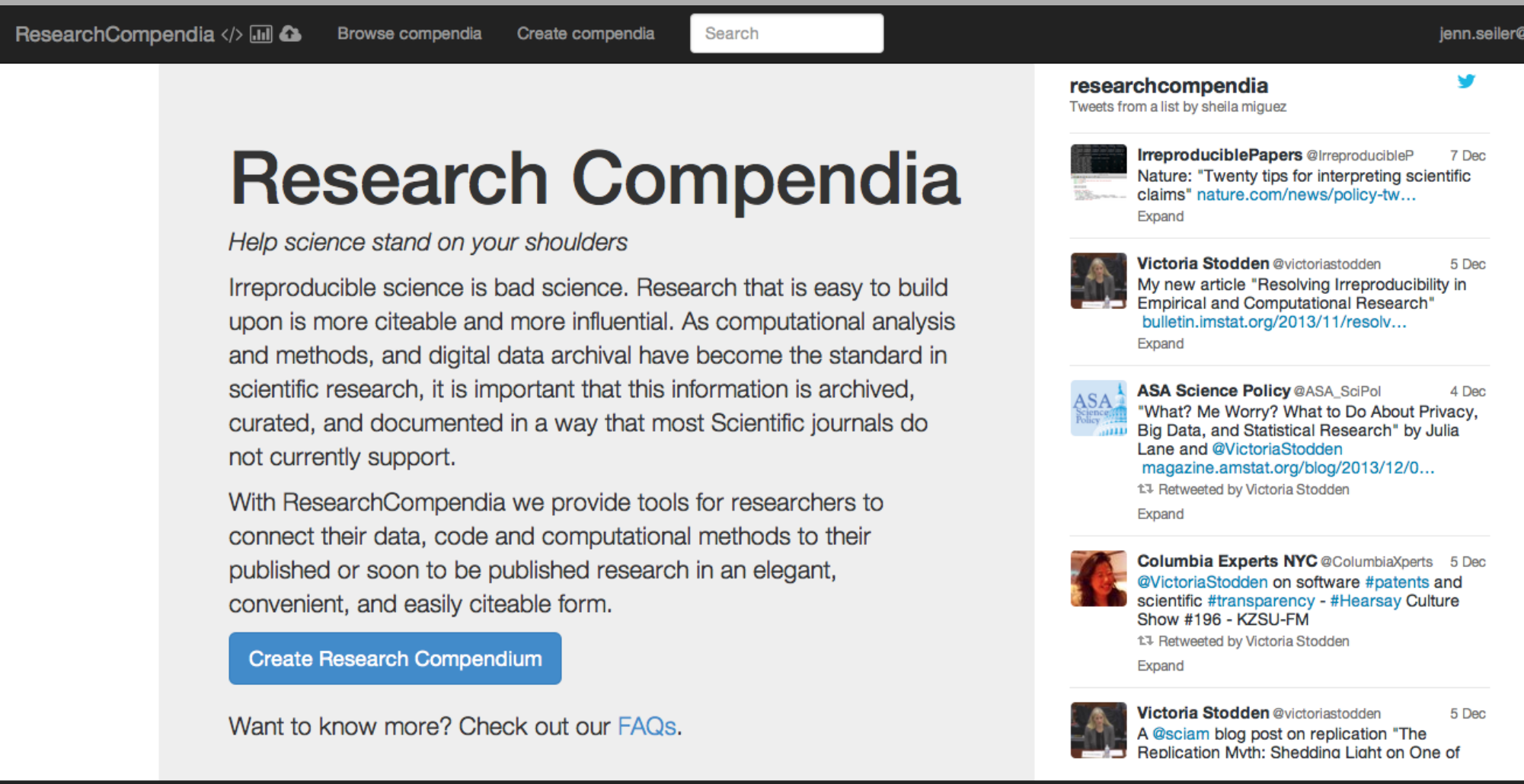
Despite the fact that the majority of papers in high impact journals to day have computational components, policies regarding codes are not advancing as quickly as code policies.

What is Research Compendia?

"We introduce the concept of a compendium as both a container for the different elements that make up the document and its computations (i.e. text, code, data,...), and as a means for distributing, managing and updating the collection." ***Gentleman, R. and Temple Lang, D. "Statistical Analyses and Reproducible Research" (May 2004). Bioconductor Project Working Papers.***

- ResearchCompendia is a web service allowing people to share the research software and data associated with a scientific publication (articles and working papers).
- We provide the tools to publish digital scholarly objects by hosting data, code, methods documentation, parameters, and environmental settings in a form that is accessible, trackable, and persistent.
- Data and code will be citable and linked to the original publication.
- Soon, we wish to support the verification and validation processes by providing for the remote execution of shared codes in our cloud resources, and the visualization of results.
- Most of all we wish to make these tools heavily automated, and easy to access and utilize to lessen the exertion required from already overburdened academic researchers in the process of publishing fully reproducible work.

ResearchCompendia: How it works



ResearchCompendia: How it works

ResearchCompendia </> Browse compendia Create compendia Search jenn.seiler@gr

10.1088/0264-93

Compendia Owner* jenn.seiler

Site user who owns this compendium

Status* draft

Title* Status of NINJA: the Numerical INJection Analysis project

Please title your compendium. Does not have to match the title of the paper.

Authors* Laura Cadonati, Benjamin Aylott, John G Baker, William D Boggs, Michael Boyle, Patrick R Brady, Duncan A Brown, Bernd Brügmann, Luisa T Buchman, Alessandra Buonanno, Jordan Camp, Manuela Campanelli, Joan Centrella, Shourov Chatterji, Nelson Christensen, Tony Chu, Peter Diener, Nils Dorband, Zachariah B Etienne, Joshua Faber, Stephen Fairhurst, Benjamin Farr, Sebastian Fischetti, Gianluca Guidi, Lisa M Goggin, Mark Hannam, Frank Herrmann, Ian Hinder, Sascha Husa, Vicky Kalogera, Drew Keppel, Lawrence E Kidder, Bernard J Kelly, Badri Krishnan, Pablo Laguna, Carlos O Lousto, Ilya Mandel, Pedro Marronetti, Richard Matzner, Sean T McWilliams, Keith D Matthews, R Adam Mercer, Satyanarayan R P Mohapatra, Abdul H Mroué, Hiroyuki Nakano, Evan Ochsner, Yi Pan, Larne Pekowsky, Harald P Pfeiffer, Denis Pollney, Frans Pretorius, Vivien Raymond, Christian Reisswig, Luciano Rezzolla, Oliver Rinne, Craig Robinson, Christian Röver, Lucía Santamaría, Bangalore Sathyaprakash, Mark A Scheel, Erik Schnetter, Jennifer Seiler, Stuart L Shapiro, Deirdre Shoemaker,

Authors listed in paper (max length 500)

Article URL http://stacks.iop.org/0264-9381/26/i=11/a=114008?key=crossref.614a353865bc379d47880dfd7e600a62

DOI 10.1088/0264-9381/26/11/114008

ResearchCompendia About FAQs Contact Terms of Use Partners Developers

Research Compendia: How it works

ResearchCompendia </> Browse compendia Create compendia Search jenn.seiler@gm

Code archive file Choose File no file selected
File containing an archive of the code. Please include a README in the archive according to site recommendations.

Data archive file Choose File no file selected
File containing an archive of the data. Please include a README in the archive according to site recommendations.

Article file Choose File no file selected
File containing the article. Optional.

Content license ---------

Code license ---------

Compendium type ---------

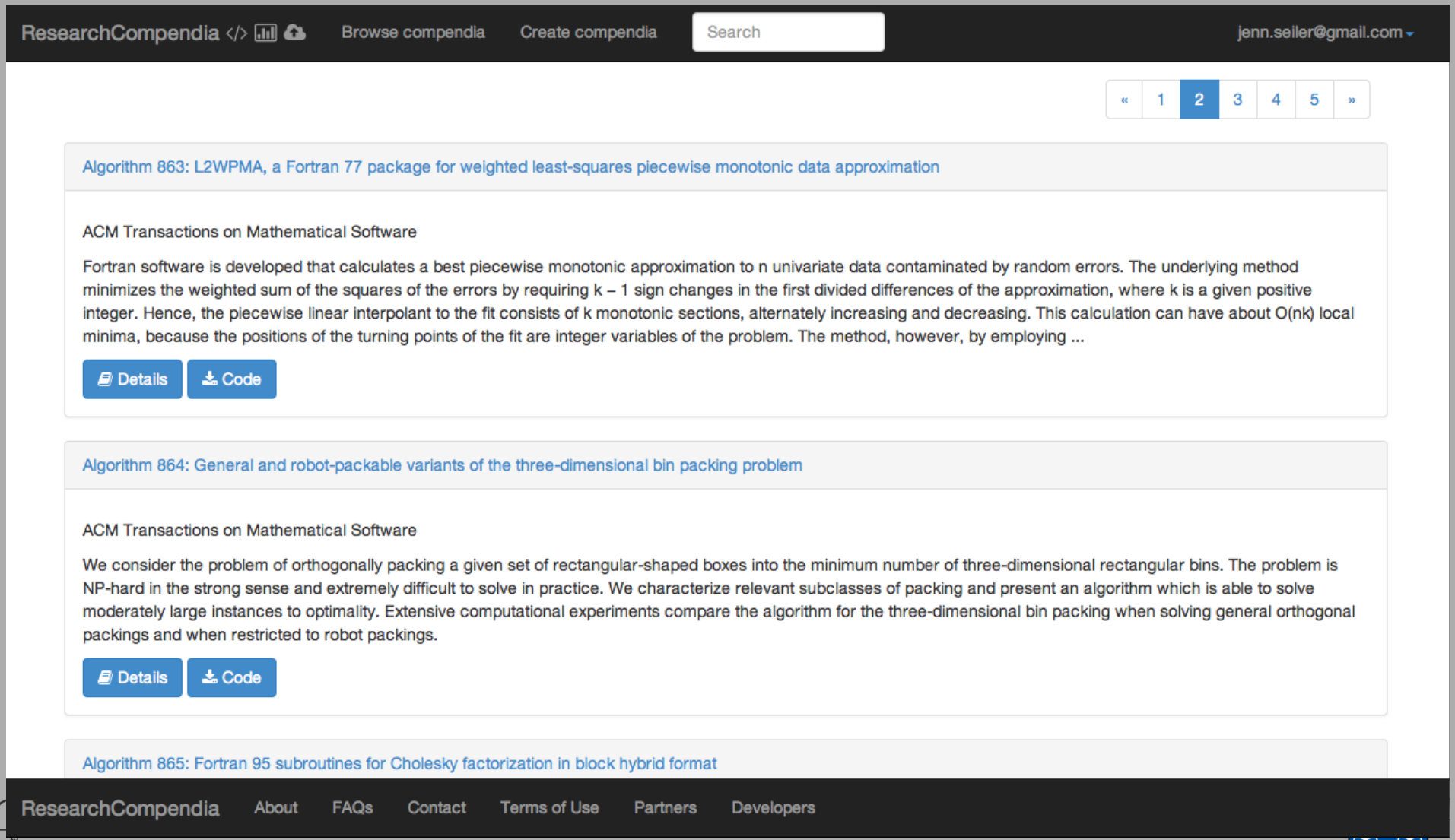
Primary research field ---------

Secondary research field ---------

Notes for staff

ResearchCompendia About FAQs Contact Terms of Use Partners Developers

Research Compendia: How it works



Research Compendia: How it works

ResearchCompendia </> Browse compendia Create compendia Search jenn.seiler@

Are Public Investment Efficient in Creating Capital Stocks in Developing Countries? Estimates of Government Net Capital Stocks for 26 Developing Countries, 1970-2002

Christophe Hurlin, Florence Arestoff Coders: Christophe Hurlin, Florence Arestoff

Code and Data Abstract

This code computes various estimates of the government net capital stocks for a panel of 26 developing countries over the period 1970-2001. The authors can choose the efficiency parameter (i.e. the percentage of public investments that are really used to create new capital stock). The Perpetual Inventory Method (PIM) corresponds to a case with an efficiency parameter equal to one. The data of public capital stocks can be expressed in Local Currency Unit or in percentage of real GDP. For constant prices, the base year is not the same for all the countries and corresponds to the base year used for the GDP implicit price (WDI Code: Y.GDP.DEFL.ZS). See WDI (2004) and Hurlin and Arestoff (2004), pages 7-8. The data can be downloaded in a csv format (Sheet "Results").

</> code data

Paper Abstract

We provide various estimates of the government net capital stocks for a panel of 26 developing countries over the period 1970-2001. These internationally comparable series of public capital are proposed as a complementary solution to the use of public investment flows and to the use of physical measures of infrastructure when one comes to evaluate the productivity of the public capital formation in developing countries. In these estimates based on various assumptions, we attempt to take into account the potential inefficiency of public investment in creating capital.

Page Owner

sheila@codersquid.com

created 11/11/2013

ResearchCompendia About FAQs Contact Terms of Use Partners Developers

Research Compendia: Features

To do this we hope to provide the following tools:

- Quick and elegant Compendia Page creation with easy to navigate access to all relevant data, code, documentation, and results, with no coding required by the researcher. [current]
- Free data and code hosting. [current]
- Trackable code and data usage monitoring and version tracking. [current]
- Highly visible and easily citable code, with DOI issuing for data and code objects. [soon]
- Executable functionality with easy parameter entry that enables users and contributors alike to run compendia codes in our cloud and obtain requested results in an downloadable file with optional results visualizations for some languages (R, MatLab, Python, Cactus, etc.). [planned]

Research Compendia: Goals

ResearchCompendia follows these main objectives:

- To allow researchers to quickly disseminate internationally the results of their research, which will considerably increase the potential of citations for their papers.
- To provide a very large community of users with the ability to use the latest scientific methods in a user-friendly environment. This will speed up the process of converting scientific results into productive forces.
- To allow members of the academic community (researchers, editors, referees, etc.) to replicate scientific results and to demonstrate their robustness.
- To provide a forum for the discussion/execution of research verification and communication.

Research Compendia: Collaboration

ResearchCompendia </> Browse compendia Create compendia Search jenn.s Googl

Partners

Fellows

Funders

Journals

Institutions

The ResearchCompendia team is extremely grateful to all of our partners for their support, advice, and assistance.

ALFRED P. SLOAN FOUNDATION

rackspace®
the open cloud company

COLUMBIA UNIVERSITY
ACADEMIC COMMONS

Want to help?

ResearchCompendia About FAQs Contact Terms of Use Partners Developers

Development: Open Source for Open Science

Fork me on GitHub

Development

ResearchCompendia is a free and open source software project. If you are interested in contributing to a project that promotes reproducible and open science, please talk to us! This is a project to allow scientists to create compendium comprising all relevant narrative, code, and data to make their research truly reproducible. Our goal is allow and teach researchers to document the computational portions of their research methods as thoroughly as they would document a tabletop experiment. We want our tools to fulfill these goals:

- We will make it possible to archive all of the data, codes, documentation, parameters, and environmental settings linked with published research in a versioned form.
- We will support the verification and validation processes by providing for the execution of shared code and the visualization of results.
- We want to help and encourage researchers to manage their research in a way that makes it remixable and executable.
- Most of all we wish to make these tools heavily automated, and easy to access and utilize to lessen the exertion required from already overburdened academic researchers in the process of publishing fully reproducible work.

Imagine if all the materials in a research project could be continuously packaged and deployed with no snags preventing use and refinement by anyone. We could help make research accessible to everyone.

Visit our our github repo to check out the source code that runs this site. Read our docs on contributing. We maintain an list of issues with labels to guide contributions, We welcome participation from people of all skill levels and backgrounds. You don't need to be a web developer or programmer. We welcome people who are interested in testing, documentation, design, and brainstorming. We are happy to hear what feature additions people would like to have and build, such as further supported packages and languages.

- Our technical documentation lives at http://tyler.rtfd.org.
- We use the issue tracker on github. https://github.com/researchcompendia/tyler/issues
- Sheila idles in our IRC channel as *skay*, #hackingscience on freenode.net, and is occasionally joined by the travis-ci bot (hopefully with good news about the build).
- Beta site: http://researchcompendia.org
- Test site: http://labs.researchcompendia.org

Concluding Remarks

Built tools to:

- deposition/curation of versioned data and code,
- link to published article,
- provide for permanence of link.

Opportunities:

- culture change regarding digital scholarly objects
- tool development facilitating shareable objects

Challenges:

- defining ‘digital scholarly object’
- providing appropriate sharing modalities, ie for code
- intellectual property law
- partnerships across academic silos, incentives to share

Thank you.

References

A Abbott.
Disputed results a fresh blow for social psychology.
Nature, 497(7447):16, 2013.

Florian Prinz, Thomas Schlange, and Khusru Asadullah.
Believe it or not: how much can we rely on published data on potential drug targets?
Nature reviews Drug discovery, 10(9):712–712, 2011.

C Glenn Begley and Lee M Ellis.
Drug development: Raise standards for preclinical cancer research.
Nature, 483(7391):531–533, 2012.

Carl Zimmer.
A sharp rise in retractions prompts calls for reform.
The New York Times, 2012.

Aaron Mobley, Suzanne K Linder, Russell Braeuer, Lee M Ellis, and Leonard Zwelling.
A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic.
PloS one, 8(5):e63221, 2013.

Chris Drummond.
Replicability is not reproducibility: Nor is it good science, June 2009.

Victoria Stodden, Jonathan Borwein, and David H Bailey.
Òsetting the default to reproducibleÓ in computational science research.

David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, et al.
Life in the network: the coming age of computational social science.
Science (New York, NY), 323(5915):721, 2009.

Anthony JG Hey and AE Trefethen.
The data deluge: An e-science perspective.
2003.

SINTEF.
Big data, for better or worse: 90% of world's data generated over last two years.
ScienceDaily, May 2013.

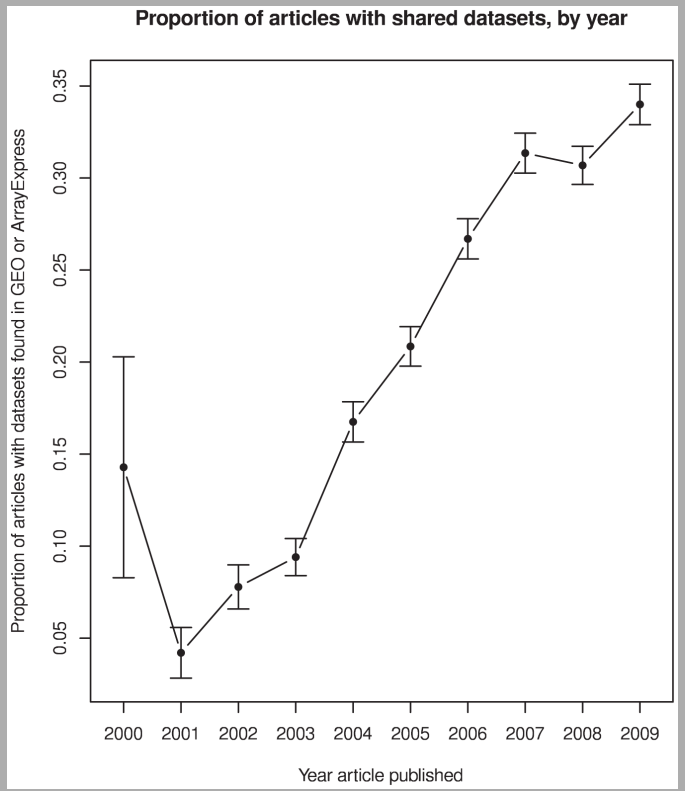
Data Sharing Resources

Lists:

re3data re3data.org (623)
DataBib databib.org (602)

Individual:

ResearchCompendia ResearchCompendia.org
FigShare figshare.com
Dataverse thedata.harvard.edu
NSCID ncsid.org
NeuroMorpho NeuroMorpho.org
OceanDataPortal www.oceandataportal.org
DRYAD datadryad.org
DataHub datahub.io
Open Science Framework osf.io
...



Code Sharing Resources

17 (vs 602) listed Repositories accept code:

ResearchCompendia ResearchCompendia.org
Dataverse thedata.harvard.edu
NanoHub nanohub.org
CRAN cran.org
MLOSS mloss.org
CLUES clues-project.org
GitHub github.com
SourceForge sourceforge.net
Launchpad launchpad.net
GoogleCode code.google.com
RunMyCode runmycode.org

Perhaps the terrible numbers for code sharing can be partly attributes to the lack of infrastructure and available tools.

Policy: Catching up to the Trend

NSF grant guidelines:

> *"NSF. . . expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages grantees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable." (2005 and earlier)*

NIH (2003):

> *"The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers."*

NAS (2003):

> *"Principle 1. Authors should include in their publications the data, algorithms, or other information that is central or integral to the publication–that is, whatever is necessary to support the major claims of the paper and would enable one skilled in the art to verify or replicate the claims."*