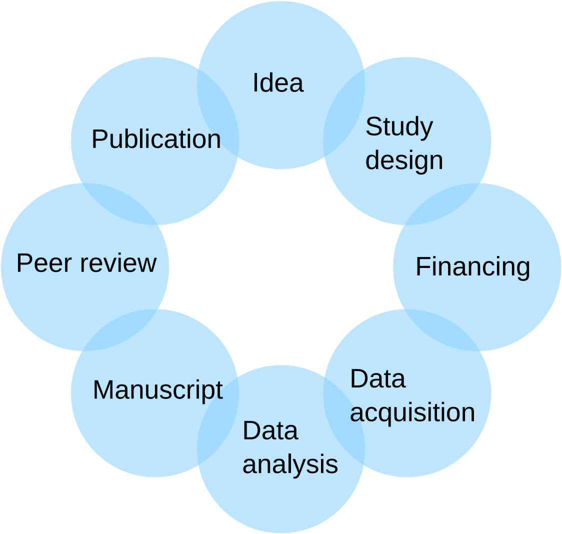
Open Science in all its Facets with a Focus on Research Software

Konrad U. Förstner

ZB MED – Information Center Life Sciences, Cologne, Germany &
TH Köln, Cologne Germany

Novel approaches in Open Science
University Library, University of Mannheim
Oct. 22nd 2018



Idea

Study design

Financing

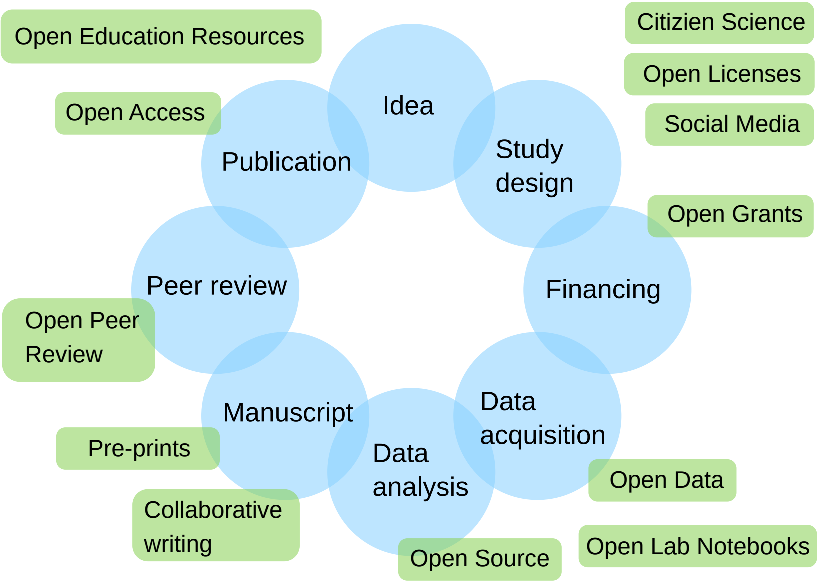
Data acquisition

Data analysis

Manuscript

Peer review

Publication


Open Education Resources
Citizien Science
Open Licenses
Social Media
Open Access
Idea
Study design
Publication
Open Grants
Peer review
Financing
Open Peer Review
Manuscript
Data acquisition
Pre-prints
Data analysis
Open Data
Collaborative writing
Open Source
Open Lab Notebooks

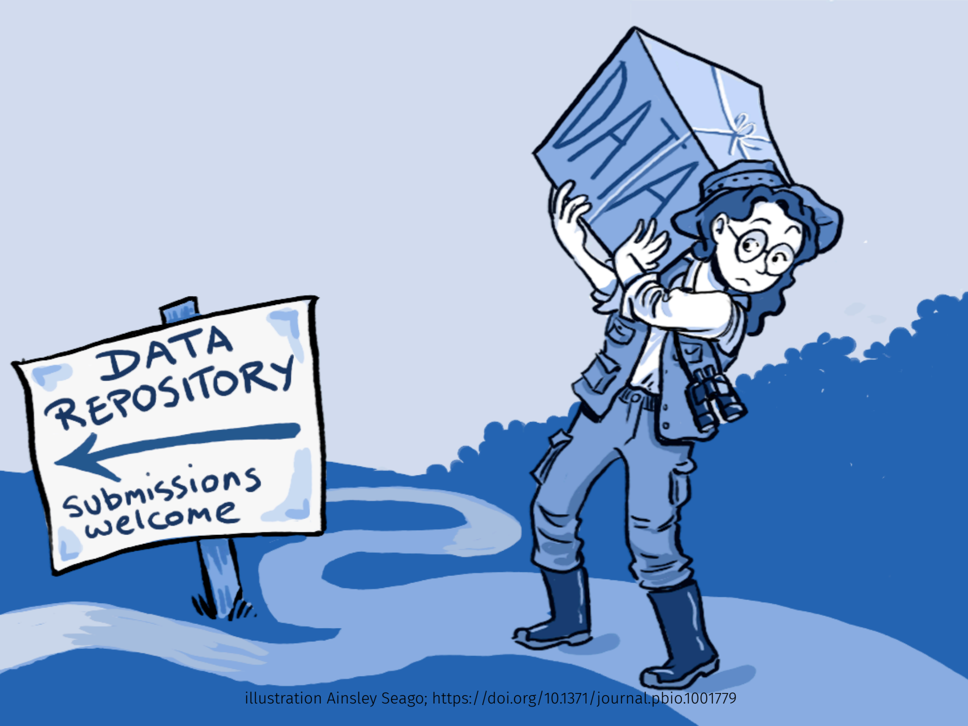
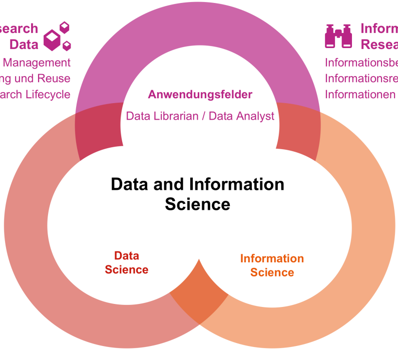


illustration Ainsley Seago; https://doi.org/10.1371/journal.pbio.1001779

Shifting to Data Savvy: The Future of Data Science In Libraries

Burton, Matt and Lyon, Liz and Erdmann, Chris and Tijerina, Bonnie (2018)
Shifting to Data Savvy: The Future of Data Science In Libraries. Project Report.
University of Pittsburgh, Pittsburgh, PA; https://d-scholarship.pitt.edu/id/eprint/33891

New Bachelor at TH Köln: *Data and Information Science*



Informationsgesellschaft

Informationsethik
Informationsrecht • Datenschutz

Research Data

Data Management
Data Sharing und Reuse
Research Lifecycle

Information Research

Informationsbewertung
Informationsrecherche
Informationen im Unternehmen

Anwendungsfelder

Data Librarian / Data Analyst

Data and Information Science

Data Science

Information Science

Digging into Data

Datenmodellierung
Datentransformation
Statistische Verfahren
Datenbanksysteme
Data- / Text- / Web-Mining
Big Data

Core Information Science

Information Retrieval
Wissensorganisation
Informationserschließung
Informationsstrukturierung
Wissensrepräsentation
Suchmaschinen und semantische Technologien

Informationssysteme

Software- und Webentwicklung
Content-Management-Systeme

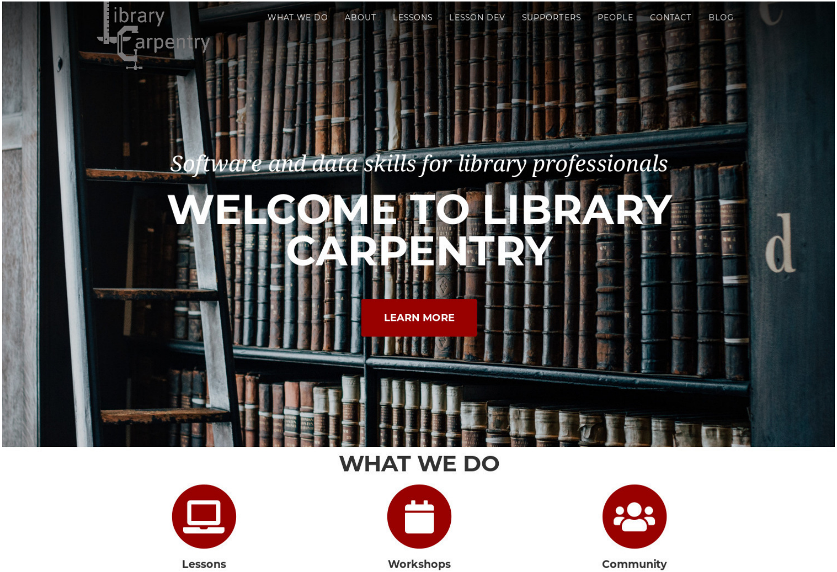
Informationsanalyse

Informationsvisualisierung
Informetrie und Szientometrie

Currently under contruction at the ZBIW of the TH Köln:

Certificate course *Data Librarian*

https://www.everystockphoto.com/photo.php?imageId=2381033 – The Library of Congress



WHAT WE DO ABOUT LESSONS LESSON DEV SUPPORTERS PEOPLE CONTACT BLOG

Software and data skills for library professionals

WELCOME TO LIBRARY CARPENTRY

LEARN MORE

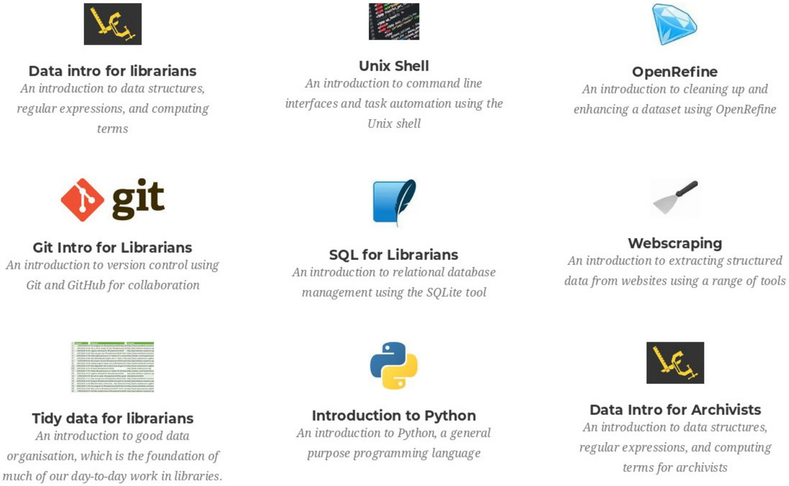
WHAT WE DO

Lessons

Workshops

Community

https://librarycarpentry.org/



Data intro for librarians
An introduction to data structures, regular expressions, and computing terms

Unix Shell
An introduction to command line interfaces and task automation using the Unix shell

OpenRefine
An introduction to cleaning up and enhancing a dataset using OpenRefine

Git Intro for Librarians
An introduction to version control using Git and GitHub for collaboration

SQL for Librarians
An introduction to relational database management using the SQLite tool

Webscraping
An introduction to extracting structured data from websites using a range of tools

Tidy data for librarians
An introduction to good data organisation, which is the foundation of much of our day-to-day work in libraries.

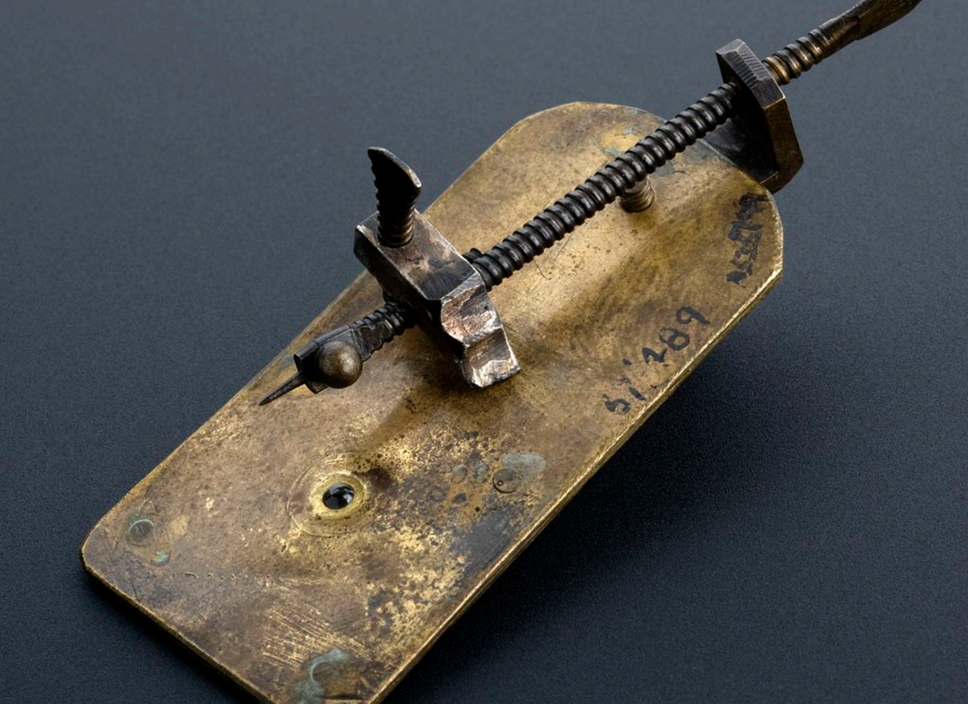
Introduction to Python
An introduction to Python, a general purpose programming language

Data Intro for Archivists
An introduction to data structures, regular expressions, and computing terms for archivists

https://librarycarpentry.org/

Science ⇌ Technology

https://commons.wikimedia.org/wiki/File:Bertini_fresco_of_Galileo_Galilei_and_Doge_of_Venice.jpg - PD

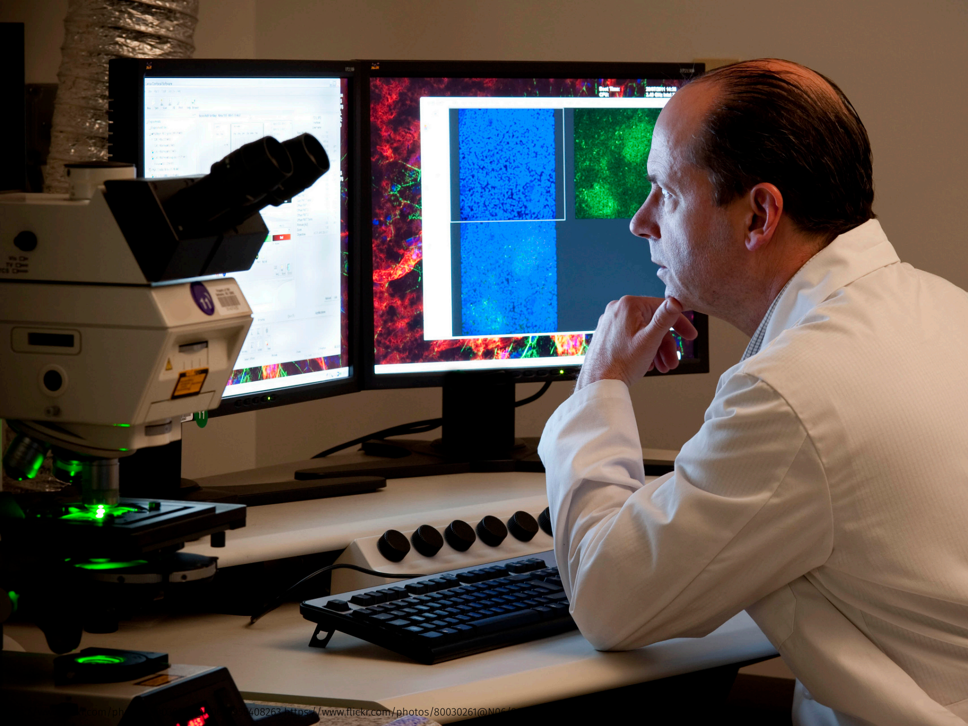


https://commons.wikimedia.org/wiki/File:Leeuwenh[illegible]simple_microscope[illegible]_Leyden,_1901-1930_Wellcome_L0057739.jpg - CC-By by Wiki Commons User Fæ

https://unsplash.com/@toddquackenbush?photo=iClZBVw5W5A - PD

Software
- an ubiquitous research tool

https://unsplash.com/@toddquackenbush?photo=IClZBVw5W5A - PD



...kr.com/ph... ...408263 https://www.flickr.com/photos/80030261@N06/...

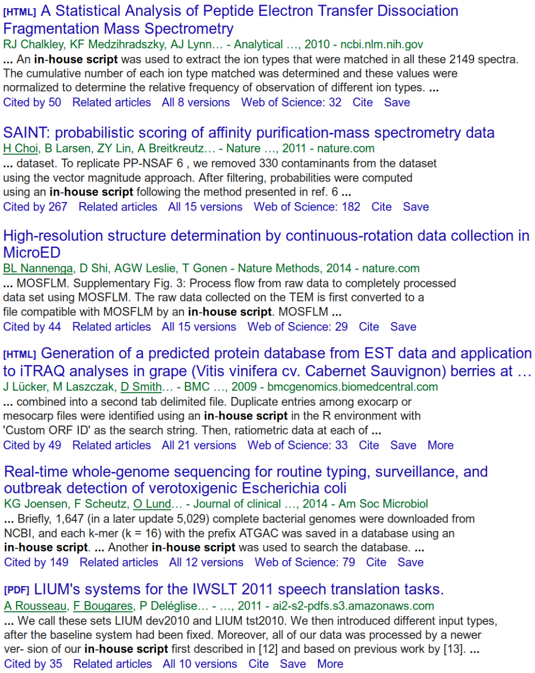
[illegible]08263 https://www.flickr.com/photos/80030261@N06/9955408263 - CC-By by Wiki Commons User EduVolunteer

It is unquestionable that there is a strong and growing dependence of research on software.

Software is also a result of the scientific work.

Quality, accessibility, citability, etc. have to be ensured.

The importance of software for research is widely ignored.



[HTML] A Statistical Analysis of Peptide Electron Transfer Dissociation Fragmentation Mass Spectrometry
RJ Chalkley, KF Medzihradszky, AJ Lynn… - Analytical …, 2010 - ncbi.nlm.nih.gov
... An **in-house script** was used to extract the ion types that were matched in all these 2149 spectra. The cumulative number of each ion type matched was determined and these values were normalized to determine the relative frequency of observation of different ion types. **...**
Cited by 50 Related articles All 8 versions Web of Science: 32 Cite Save

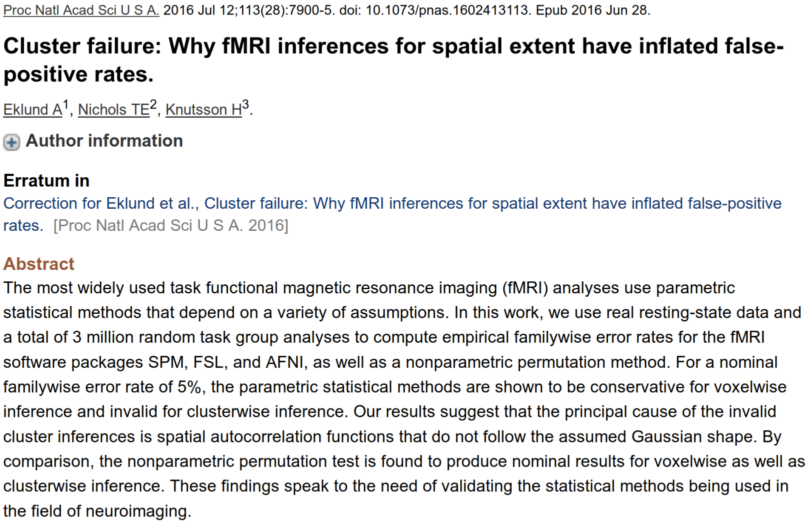
SAINT: probabilistic scoring of affinity purification-mass spectrometry data
H Choi, B Larsen, ZY Lin, A Breitkreutz… - Nature …, 2011 - nature.com
... dataset. To replicate PP-NSAF 6 , we removed 330 contaminants from the dataset using the vector magnitude approach. After filtering, probabilities were computed using an **in-house script** following the method presented in ref. 6 **...**
Cited by 267 Related articles All 15 versions Web of Science: 182 Cite Save

High-resolution structure determination by continuous-rotation data collection in MicroED
BL Nannenga, D Shi, AGW Leslie, T Gonen - Nature Methods, 2014 - nature.com
... MOSFLM. Supplementary Fig. 3: Process flow from raw data to completely processed data set using MOSFLM. The raw data collected on the TEM is first converted to a file compatible with MOSFLM by an **in-house script**. MOSFLM **...**
Cited by 44 Related articles All 15 versions Web of Science: 29 Cite Save

[HTML] Generation of a predicted protein database from EST data and application to iTRAQ analyses in grape (Vitis vinifera cv. Cabernet Sauvignon) berries at …
J Lücker, M Laszczak, D Smith… - BMC …, 2009 - bmcgenomics.biomedcentral.com
... combined into a second tab delimited file. Duplicate entries among exocarp or mesocarp files were identified using an **in-house script** in the R environment with 'Custom ORF ID' as the search string. Then, ratiometric data at each of **...**
Cited by 49 Related articles All 21 versions Web of Science: 33 Cite Save More

Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli
KG Joensen, F Scheutz, O Lund… - Journal of clinical …, 2014 - Am Soc Microbiol
... Briefly, 1,647 (in a later update 5,029) complete bacterial genomes were downloaded from NCBI, and each k-mer (k = 16) with the prefix ATGAC was saved in a database using an **in-house script**. **...** Another **in-house script** was used to search the database. **...**
Cited by 149 Related articles All 12 versions Web of Science: 79 Cite Save

[PDF] LIUM's systems for the IWSLT 2011 speech translation tasks.
A Rousseau, F Bougares, P Deléglise… - …, 2011 - ai2-s2-pdfs.s3.amazonaws.com
... We call these sets LIUM dev2010 and LIUM tst2010. We then introduced different input types, after the baseline system had been fixed. Moreover, all of our data was processed by a newer ver- sion of our **in-house script** first described in [12] and based on previous work by [13]. **...**
Cited by 35 Related articles All 10 versions Cite Save More



Proc Natl Acad Sci U S A. 2016 Jul 12;113(28):7900-5. doi: 10.1073/pnas.1602413113. Epub 2016 Jun 28.

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

Eklund A[1], Nichols TE[2], Knutsson H[3].

⊕ **Author information**

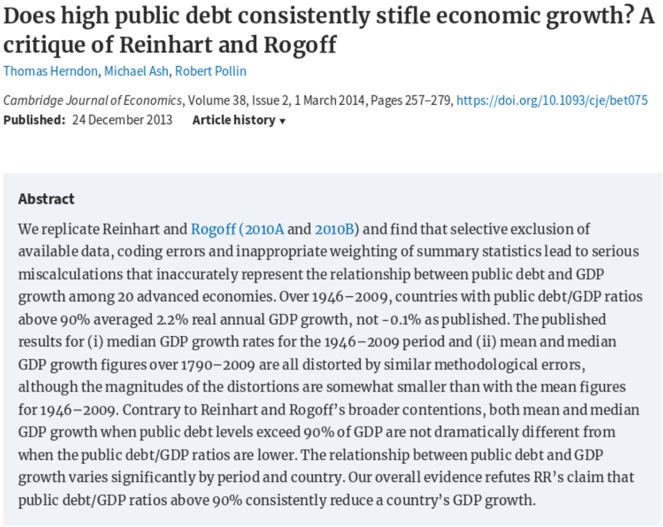
Erratum in

Correction for Eklund et al., Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. [Proc Natl Acad Sci U S A. 2016]

Abstract

The most widely used task functional magnetic resonance imaging (fMRI) analyses use parametric statistical methods that depend on a variety of assumptions. In this work, we use real resting-state data and a total of 3 million random task group analyses to compute empirical familywise error rates for the fMRI software packages SPM, FSL, and AFNI, as well as a nonparametric permutation method. For a nominal familywise error rate of 5%, the parametric statistical methods are shown to be conservative for voxelwise inference and invalid for clusterwise inference. Our results suggest that the principal cause of the invalid cluster inferences is spatial autocorrelation functions that do not follow the assumed Gaussian shape. By comparison, the nonparametric permutation test is found to produce nominal results for voxelwise as well as clusterwise inference. These findings speak to the need of validating the statistical methods being used in the field of neuroimaging.

https://doi.org/10.1073/pnas.1602413113



Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff

Thomas Herndon, Michael Ash, Robert Pollin

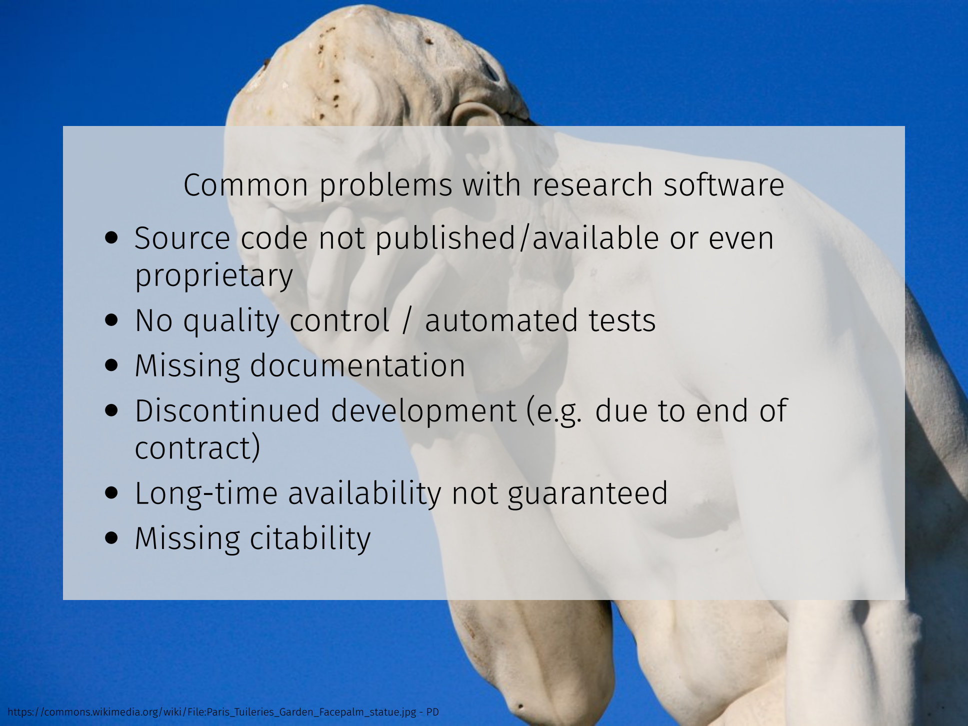
Cambridge Journal of Economics, Volume 38, Issue 2, 1 March 2014, Pages 257–279, https://doi.org/10.1093/cje/bet075

Published: 24 December 2013 **Article history** ▾

Abstract

We replicate Reinhart and Rogoff (2010A and 2010B) and find that selective exclusion of available data, coding errors and inappropriate weighting of summary statistics lead to serious miscalculations that inaccurately represent the relationship between public debt and GDP growth among 20 advanced economies. Over 1946–2009, countries with public debt/GDP ratios above 90% averaged 2.2% real annual GDP growth, not -0.1% as published. The published results for (i) median GDP growth rates for the 1946–2009 period and (ii) mean and median GDP growth figures over 1790–2009 are all distorted by similar methodological errors, although the magnitudes of the distortions are somewhat smaller than with the mean figures for 1946–2009. Contrary to Reinhart and Rogoff's broader contentions, both mean and median GDP growth when public debt levels exceed 90% of GDP are not dramatically different from when the public debt/GDP ratios are lower. The relationship between public debt and GDP growth varies significantly by period and country. Our overall evidence refutes RR's claim that public debt/GDP ratios above 90% consistently reduce a country's GDP growth.

https://doi.org/10.1093/cje/bet075



Common problems with research software

- Source code not published/available or even proprietary
- No quality control / automated tests
- Missing documentation
- Discontinued development (e.g. due to end of contract)
- Long-time availability not guaranteed
- Missing citability

https://commons.wikimedia.org/wiki/File:Paris_Tuileries_Garden_Facepalm_statue.jpg - PD

Potential reasons

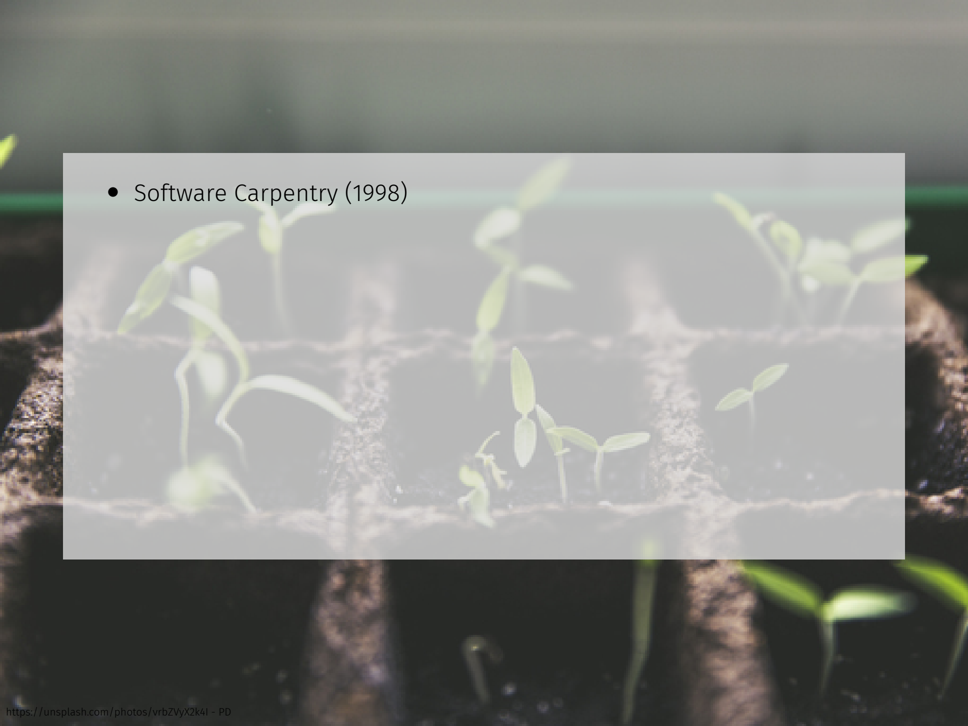
- Lack of awareness
- Lack of skills
- Lack of time
- Lack of incentives
- Lack of dedicated long-term funding
- No reviewing
- To hinder competitors

https://commons.wikimedia.org/wiki/File:Paris_Tuileries_Garden_Facepalm_statue.jpg - PD

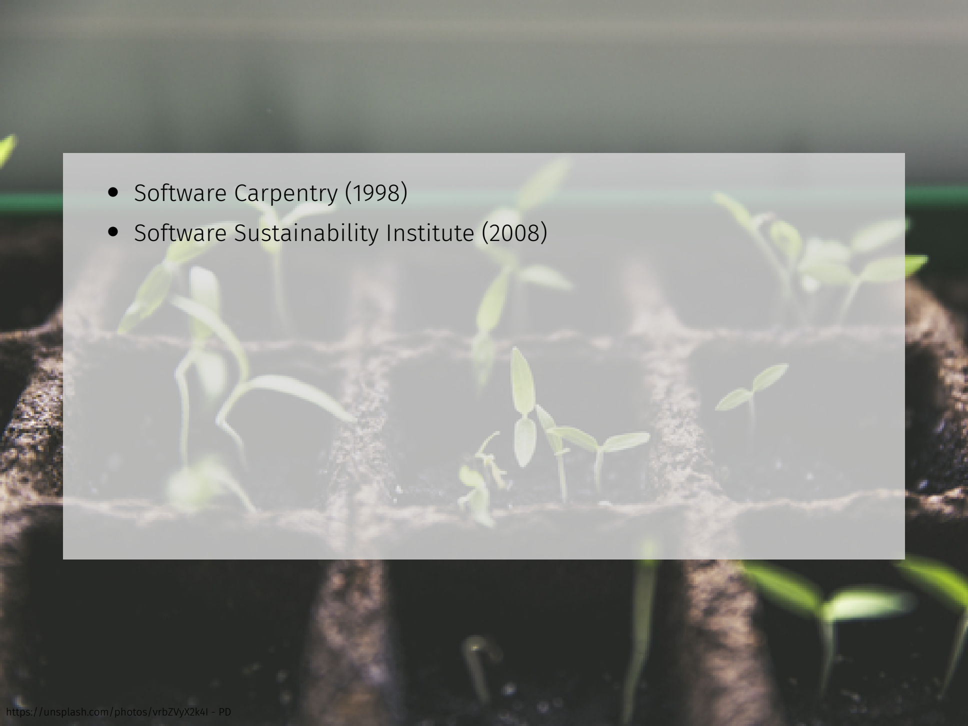
Several iniatives have been launched to address these issues.

https://unsplash.com/photos/vrbZVyX2k4I - PD

- Software Carpentry (1998)

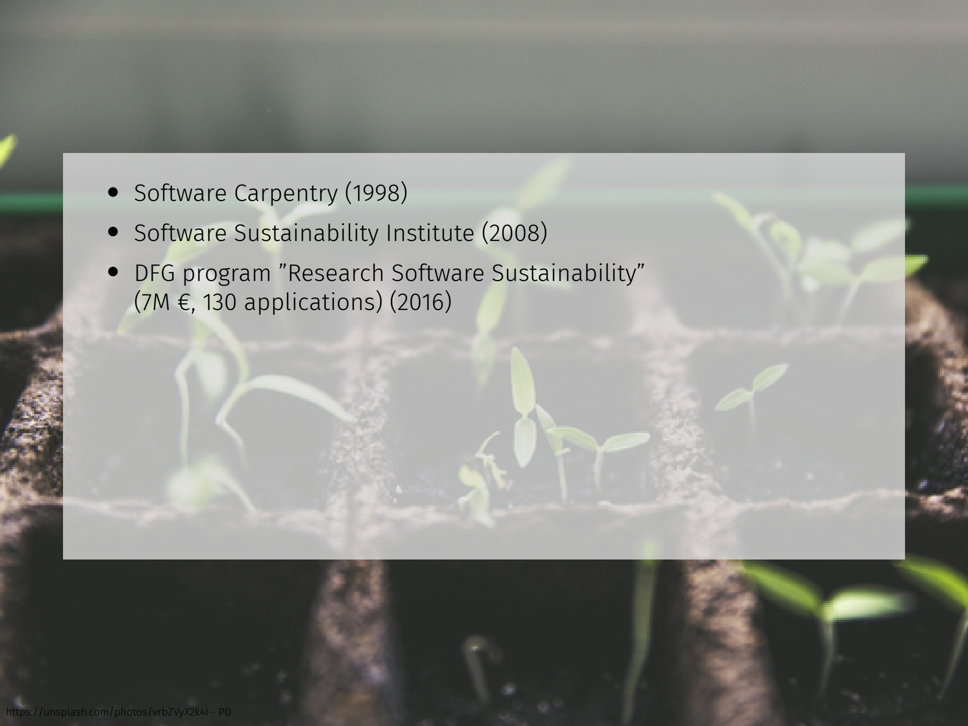


https://unsplash.com/photos/vrbZVyX2k4I - PD



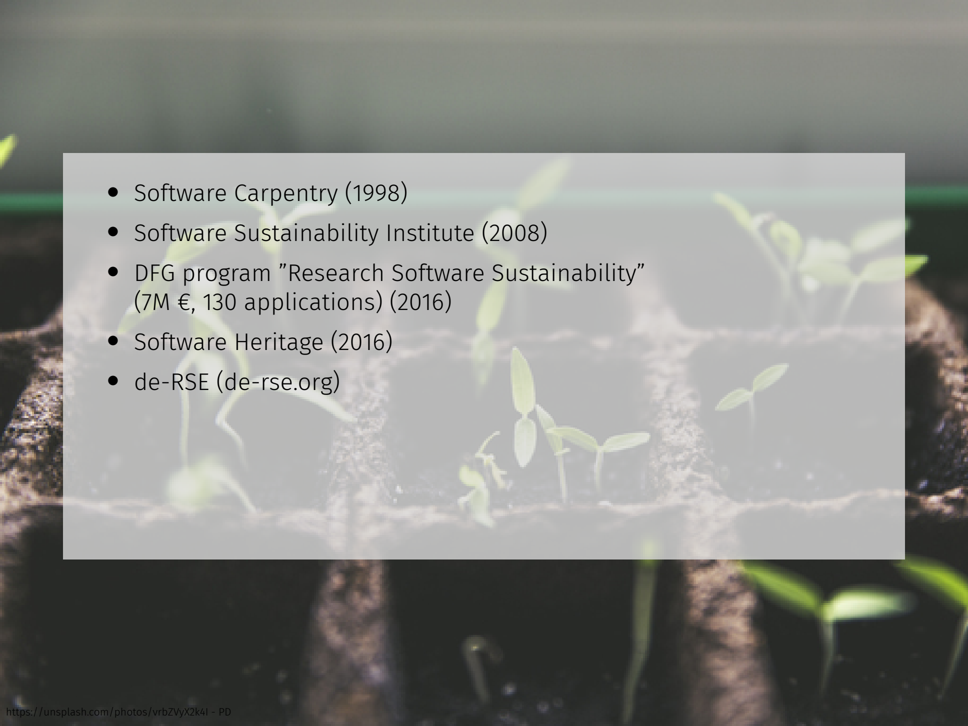
- Software Carpentry (1998)
- Software Sustainability Institute (2008)

https://unsplash.com/photos/vrbZVyX2k4I - PD



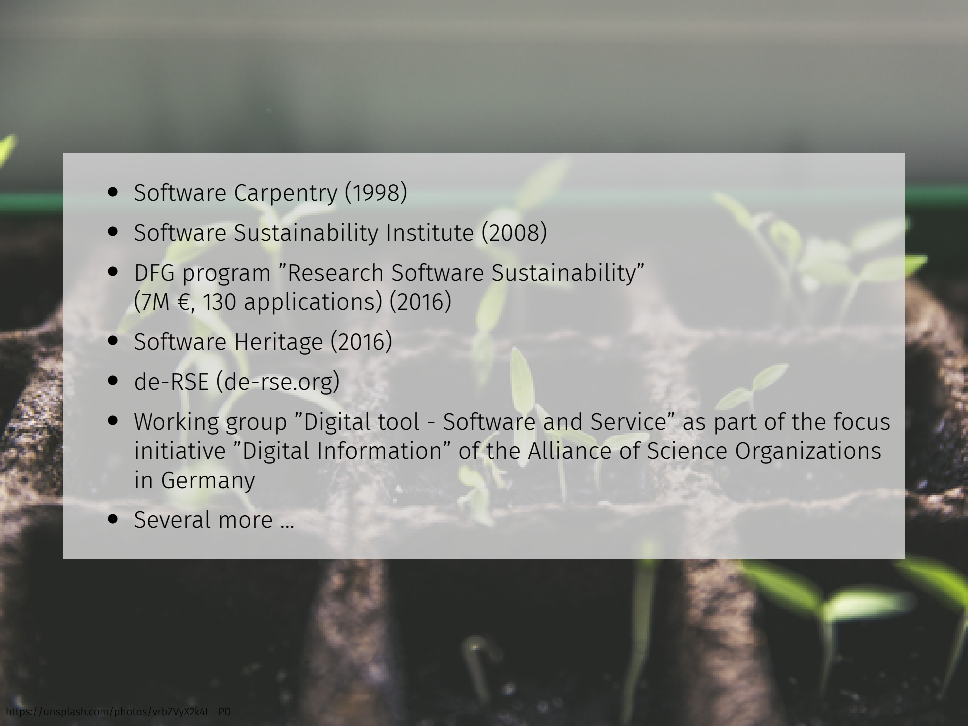
- Software Carpentry (1998)
- Software Sustainability Institute (2008)
- DFG program ”Research Software Sustainability” (7M €, 130 applications) (2016)

https://unsplash.com/photos/vrbZVyX2k4I - PD



- Software Carpentry (1998)
- Software Sustainability Institute (2008)
- DFG program ”Research Software Sustainability” (7M €, 130 applications) (2016)
- Software Heritage (2016)
- de-RSE (de-rse.org)

https://unsplash.com/photos/vrbZVyX2k4I - PD



- Software Carpentry (1998)
- Software Sustainability Institute (2008)
- DFG program ”Research Software Sustainability” (7M €, 130 applications) (2016)
- Software Heritage (2016)
- de-RSE (de-rse.org)
- Working group ”Digital tool - Software and Service” as part of the focus initiative ”Digital Information” of the Alliance of Science Organizations in Germany
- Several more ...

https://unsplash.com/photos/vrbZVyX2k4I - PD

Recommendations on the

Development, Use and Provision of Research Software

Research Software Working Group
in the Priority Initiative Digital Information
of the Alliance of German Science Organisations

https://doi.org/10.5281/zenodo.1172988

Guiding principle

The concept of Good Scientific Practice (GSP) must be also applied to research software.

But what can Good Scientific Practice mean for research software?

FAIR principle should also be applied to software

- Findable
- Accessible
- Interoperable
- Re-usable

Open

- Proper OSI conform licence

Three types of software

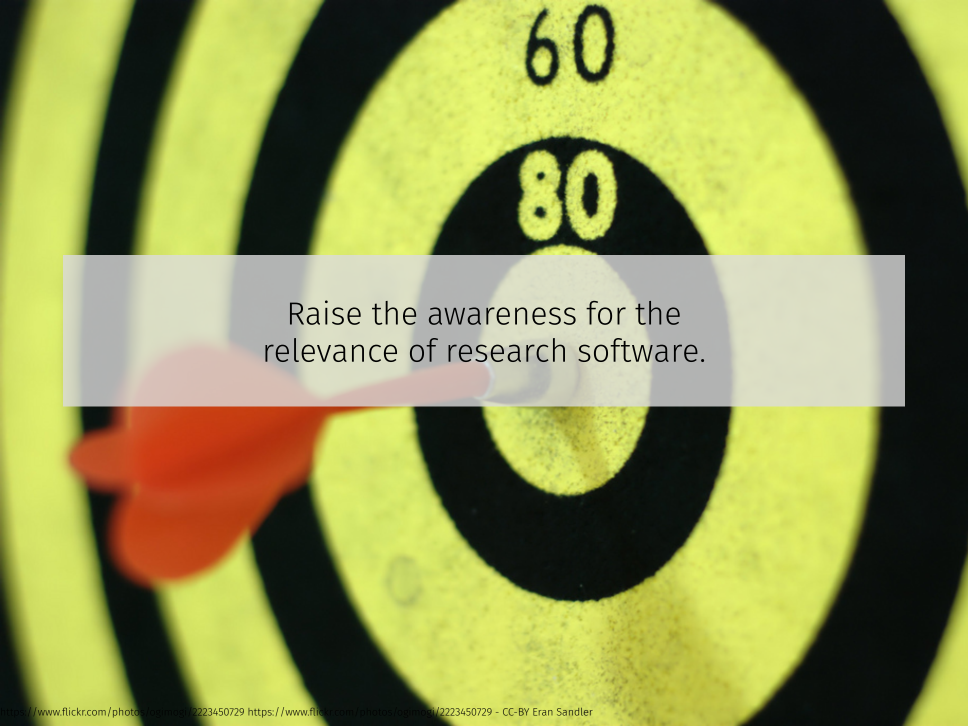
1. Small tools written for single purpose
2. Software applications (as research output)
3. Infrastructure and online services

All three levels are relevant and
have to be addressed.

Exact needs and possibilities might differ between scientific communities.

Discourse must also happen
inside of these communities.

E.g. what exactly means ”reproducibility” (bit-identical compilation?) and how long would this needed to be guaranteed?



Raise the awareness for the relevance of research software.

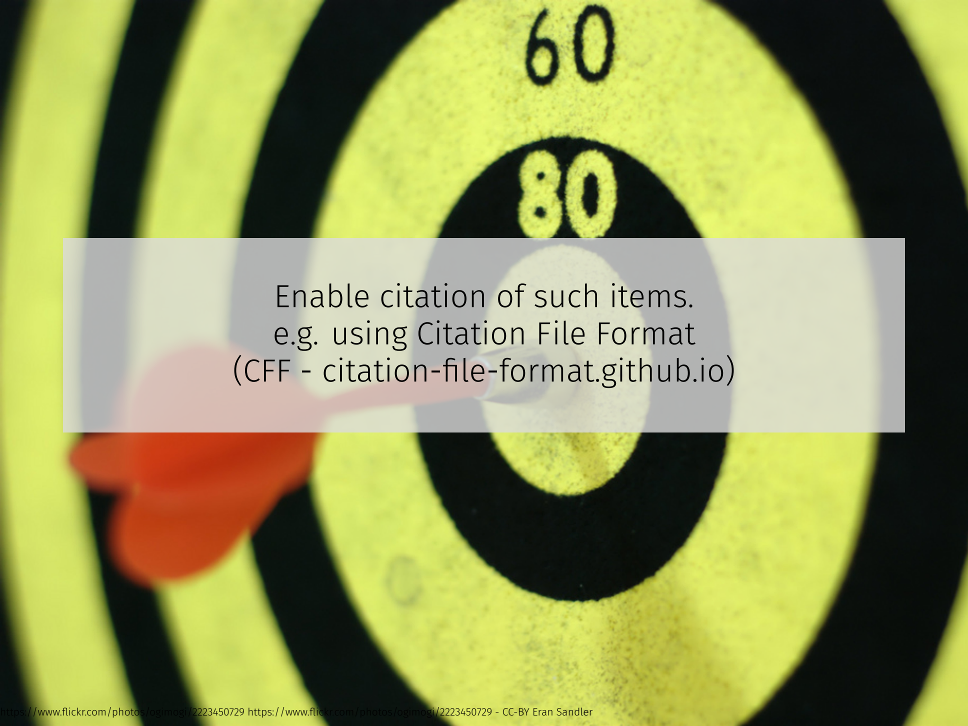
https://www.flickr.com/photos/ogimogi/2223450729 https://www.flickr.com/photos/ogimogi/2223450729 - CC-BY Eran Sandler

Introduce standards and mechanisms for quality control of research software.

https://www.flickr.com/photos/ogimogi/2223450729 https://www.flickr.com/photos/ogimogi/2223450729 - CC-BY Eran Sandler

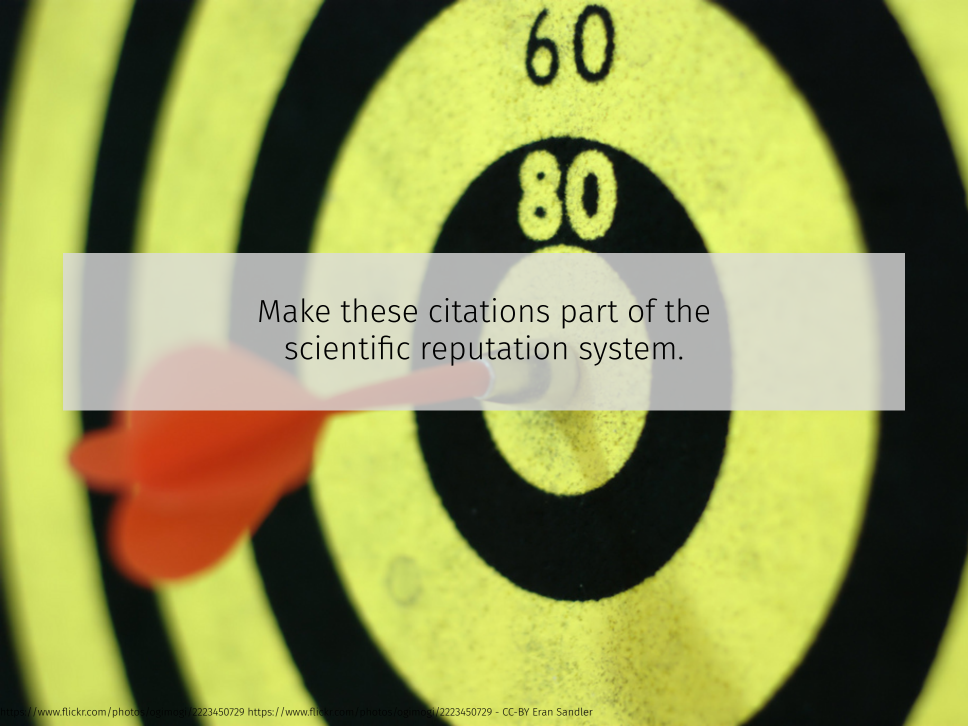
Create institutional platforms to publish and archive software/source code/workflows.

https://www.flickr.com/photos/ogimogi/2223450729 https://www.flickr.com/photos/ogimogi/2223450729 - CC-BY Eran Sandler



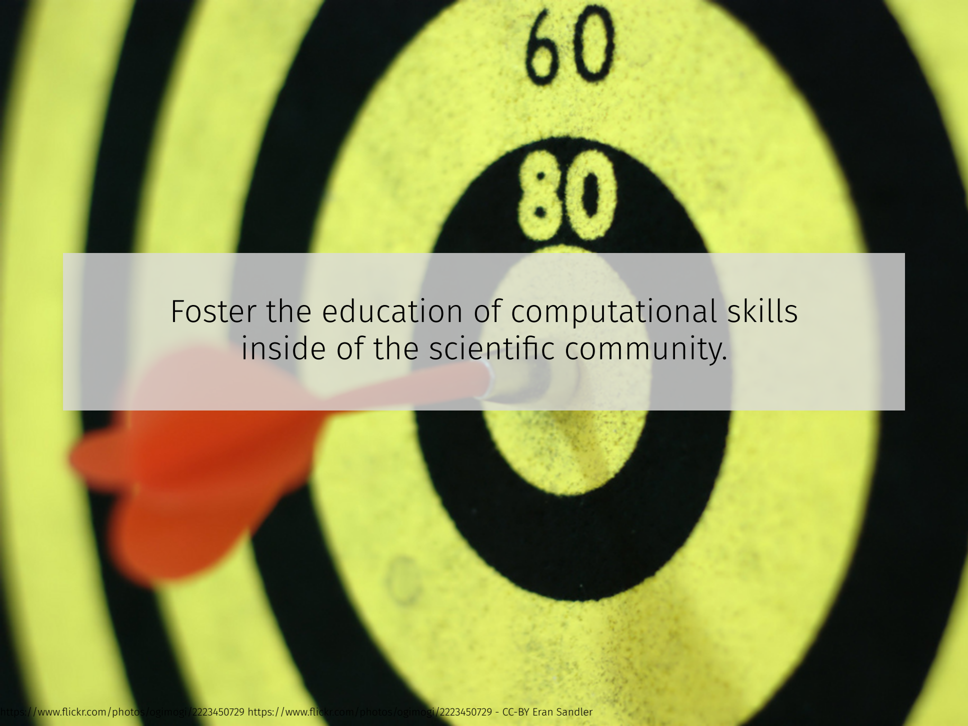
Enable citation of such items.
e.g. using Citation File Format
(CFF - citation-file-format.github.io)

https://www.flickr.com/photos/ogimogi/2223450729 https://www.flickr.com/photos/ogimogi/2223450729 - CC-BY Eran Sandler



Make these citations part of the scientific reputation system.

https://www.flickr.com/photos/ogimogi/2223450729 https://www.flickr.com/photos/ogimogi/2223450729 - CC-BY Eran Sandler

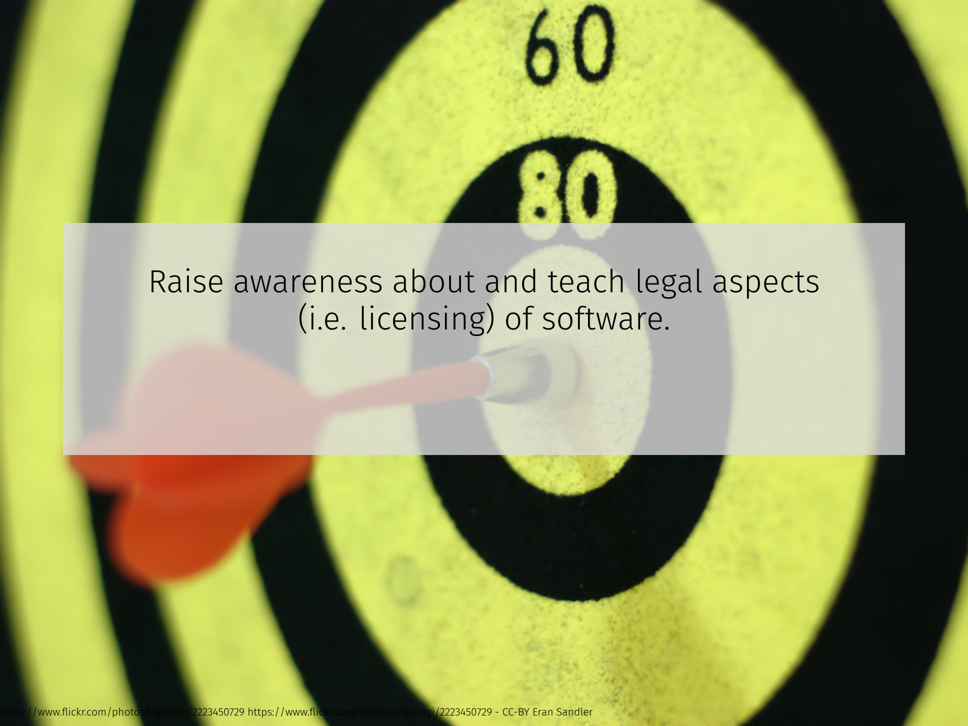


Foster the education of computational skills inside of the scientific community.

https://www.flickr.com/photos/ogimogi/2223450729 https://www.flickr.com/photos/ogimogi/2223450729 - CC-BY Eran Sandler

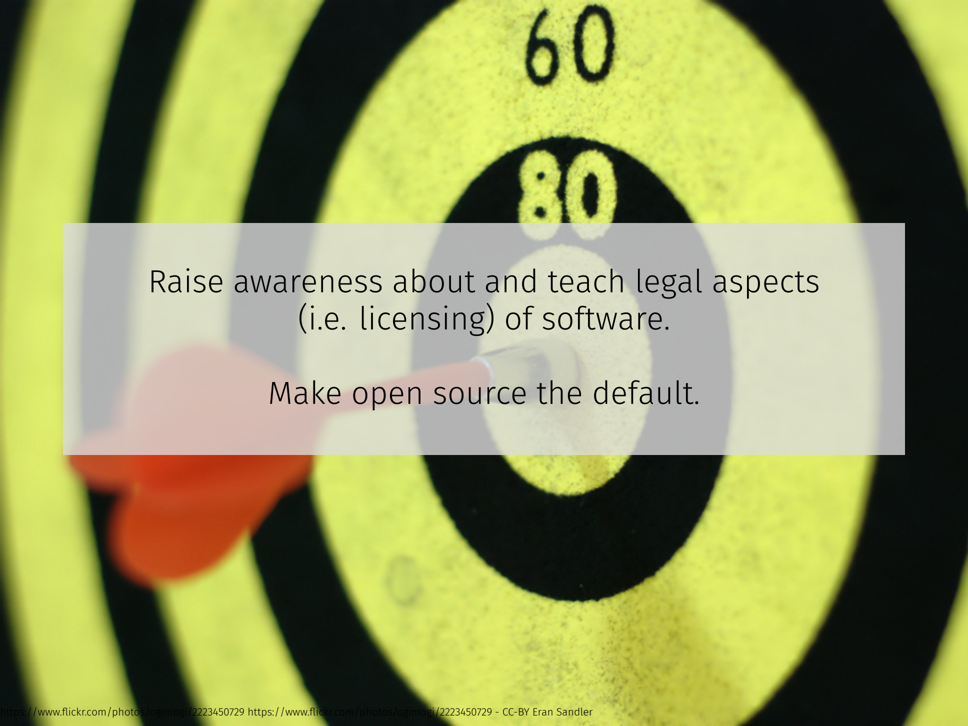
Develop new carreer paths like
Research Software Engineers,
Software Librarians, Data Scientists.

https://www.flickr.com/photos/ogimogi/2223450729 https://www.flickr.com/photos/ogimogi/2223450729 - CC-BY Eran Sandler



Raise awareness about and teach legal aspects (i.e. licensing) of software.

https://www.flickr.com/photos/ogimogi/2223450729 https://www.flickr.com/photos/ogimogi/2223450729 - CC-BY Eran Sandler



Raise awareness about and teach legal aspects (i.e. licensing) of software.

Make open source the default.

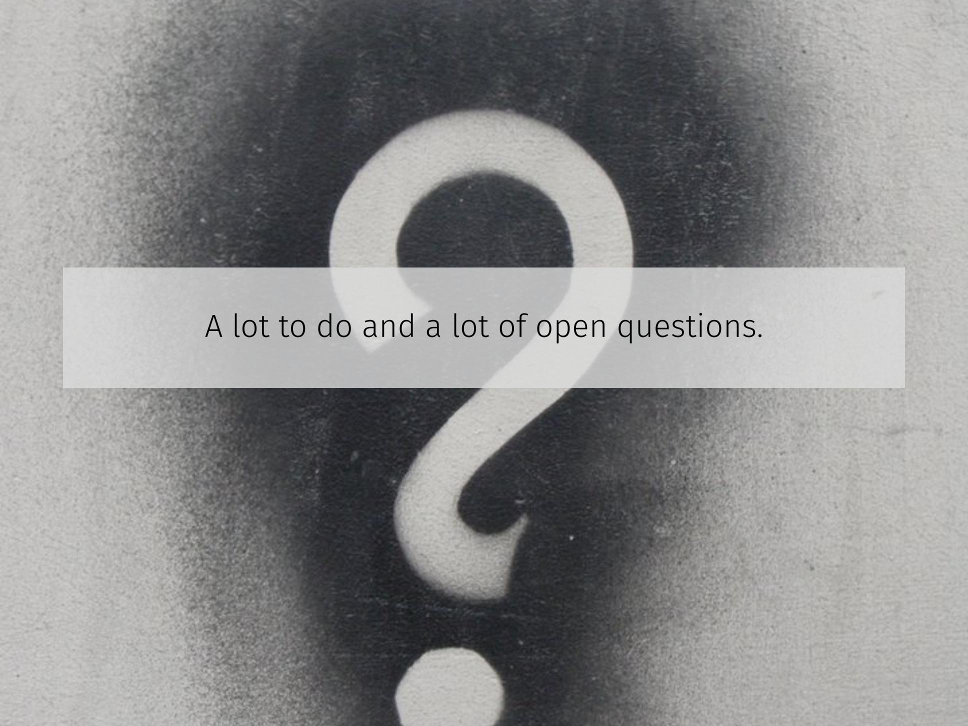
https://www.flickr.com/photos/ogimogi/2223450729 https://www.flickr.com/photos/ogimogi/2223450729 - CC-BY Eran Sandler

Facilitate the transition from single-purpose solutions to application to services.

https://www.flickr.com/photos/ogimogi/2223450729 https://www.flickr.com/photos/ogimogi/2223450729 - CC-BY Eran Sandler

Provide long-term funding to enable sustainable software development.

https://www.flickr.com/photos/ogimogi/2223450729 https://www.flickr.com/photos/ogimogi/2223450729 - CC-BY Eran Sandler



A lot to do and a lot of open questions.

www.allianzinitiative.de

Matthias Katerbow - Deutsche Forschungsgemeinschaft
Michael Goedicke - Deutsche Forschungsgemeinschaft
Leander Seige - Deutsche Forschungsgemeinschaft
Zeki Mustafa Dogan - Deutsche Forschungsgemeinschaft
Dirk Eisengräber-Pabst - Fraunhofer-Gesellschaft
Uwe Konrad - Helmholtz-Gemeinschaft
Bernadette Fritzsch - Helmholtz-Gemeinschaft
Björn Brembs - Hochschulrektorenkonferenz
Klaus Wannemacher - Hochschulrektorenkonferenz
Thomas Dandekar - Hochschulrektorenkonferenz
Georg Feulner - Leibniz-Gemeinschaft
Jürgen Fuhrmann - Leibniz-Gemeinschaft
Michael Franke - Max-Planck-Gesellschaft
Stefan Janosch - Max-Planck-Gesellschaft
Johannes Reetz -Max-Planck-Gesellschaft
Thomas Rode - Leopoldina
Mathias Bornschein - Bibliotheken der Ressortforschungseinrichtungen des Bundes

http://www.flickr.com/photos/nateone/3768979925/ – CC-BY by flickr user nateone

ZB MED

TH Köln

@konradfoerstner

http://www.flickr.com/photos/nateone/3768979925/ – CC-BY by flick user nateone