Software Sustainability Institute

Mejor Software + Mejores Datos = Mejor Investigación

Alejandra Gonzalez-Beltran, PhD
Fellow Software Sustainability Institute 2018

CIFASIS, Rosario, 23 de abril 2019

Slides: http://doi.org/10.5281/zenodo.3250706

http://orcid.org/0000-0003-3499-8262

Alejandra Gonzalez-Beltran – Technical Lead, Software Engineering Group, Science & Technology Facilities Council

@alegonbel

agbeltran

http://orcid.org/0000-0003-3499-8262

From May 2019

Research Software Engineering & Data Science
Data Standards; Data Description; Data Provenance
Semantic Web; Linked Data; Ontology Development
Digital Research Objects Dissemination
Reproducibility (ISWC challenge)
Software Engineering and Data Science Teaching
Diversity

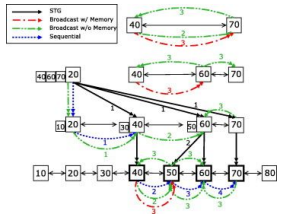
Strong Background in Maths & Computer Science

E-Government apps
Financial web apps
Code Slicing Techniques

Software Engineer

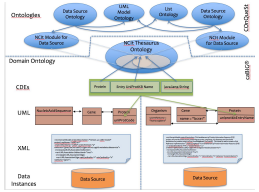
Efficient Access to Distributed Information Using Structured Peer-to-Peer Systems

Skip Tree Graph
Probabilistic Data Structure



MSc & PhD Computer Science

conquest
(cancer **on**tology **que**rying **s**ystem)



Data integration & linked data for cancer databases

isatools
http://isa-tools.org

isacreator OntoMaton linkedisa Risa isaAPI

stato statistics ontology

DATS DatA Tag Suite

OBI ontology

FAIRsharing.org standards, databases, policies

Machine-Actionable Metadata Models
JSONLDschema
JSON-schema-documenter
JSON-comparator-and-view

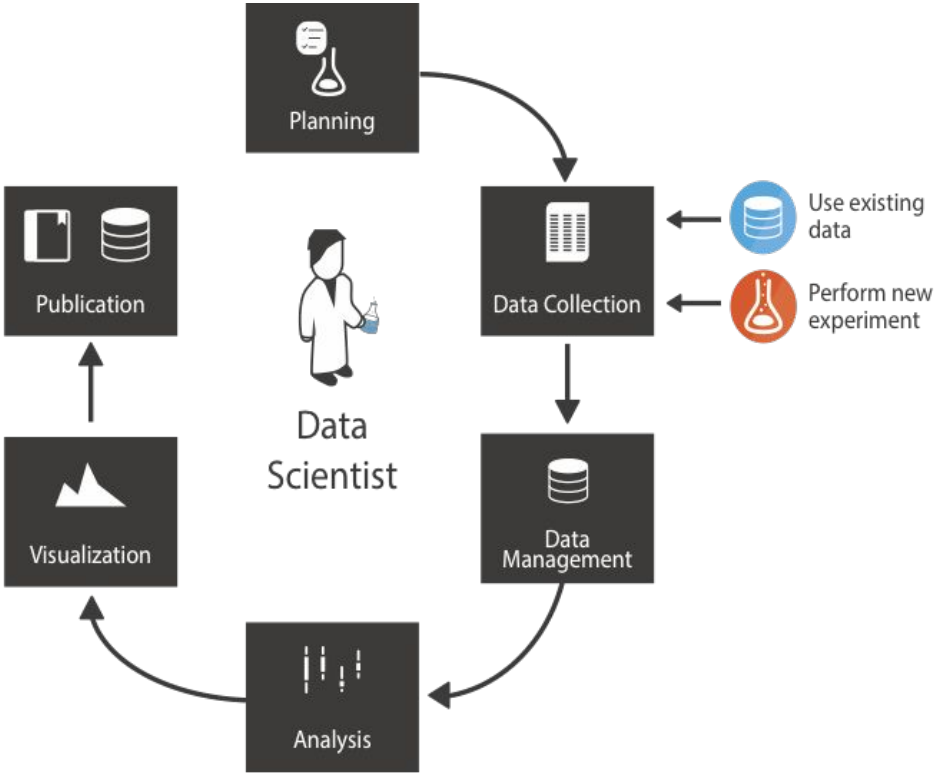
dataMED BETA version

W3C Data Catalog (DCAT) Vocabulary

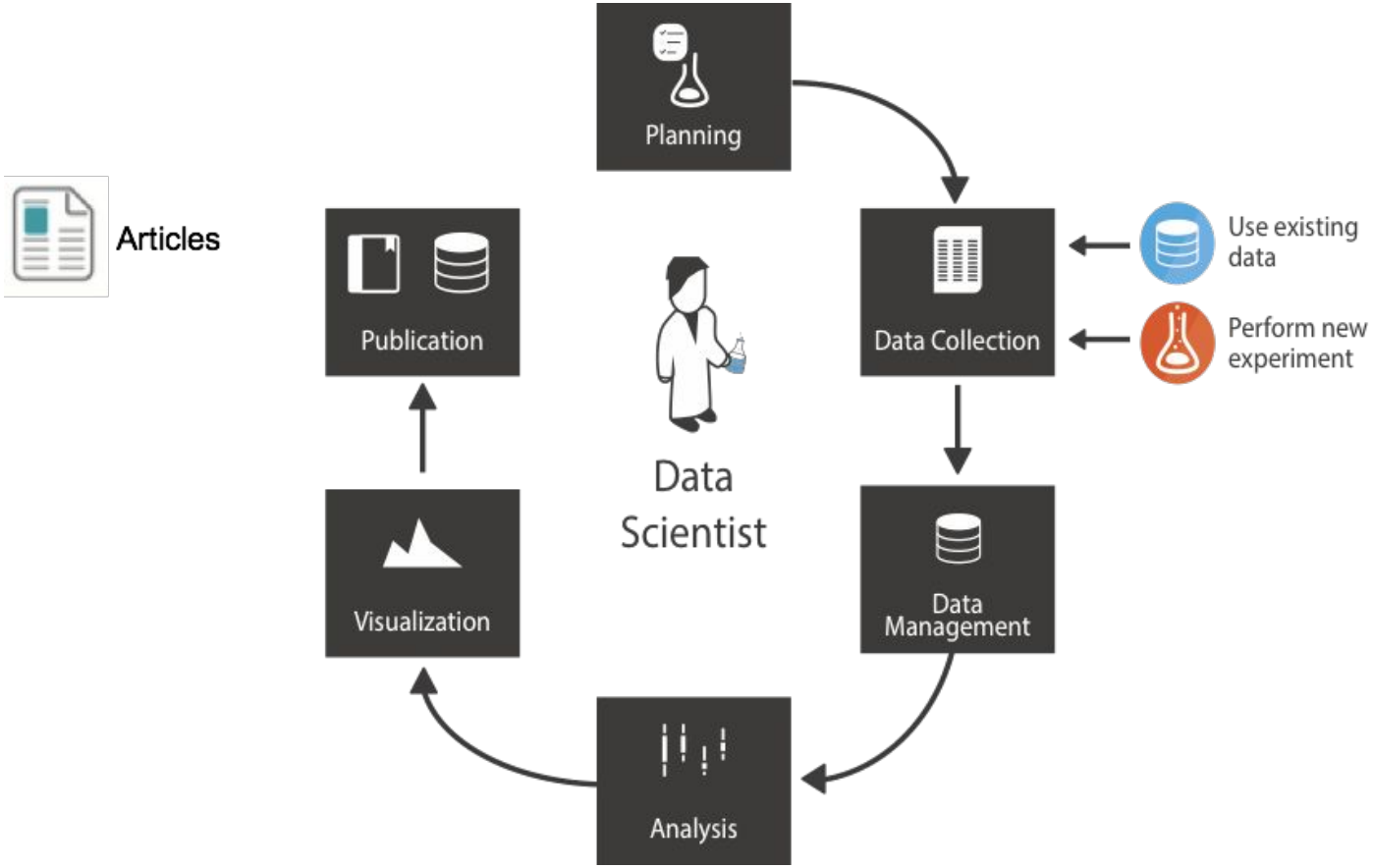
Bespoke systems to manage experimental data from the large scale scientific facilities at STFC

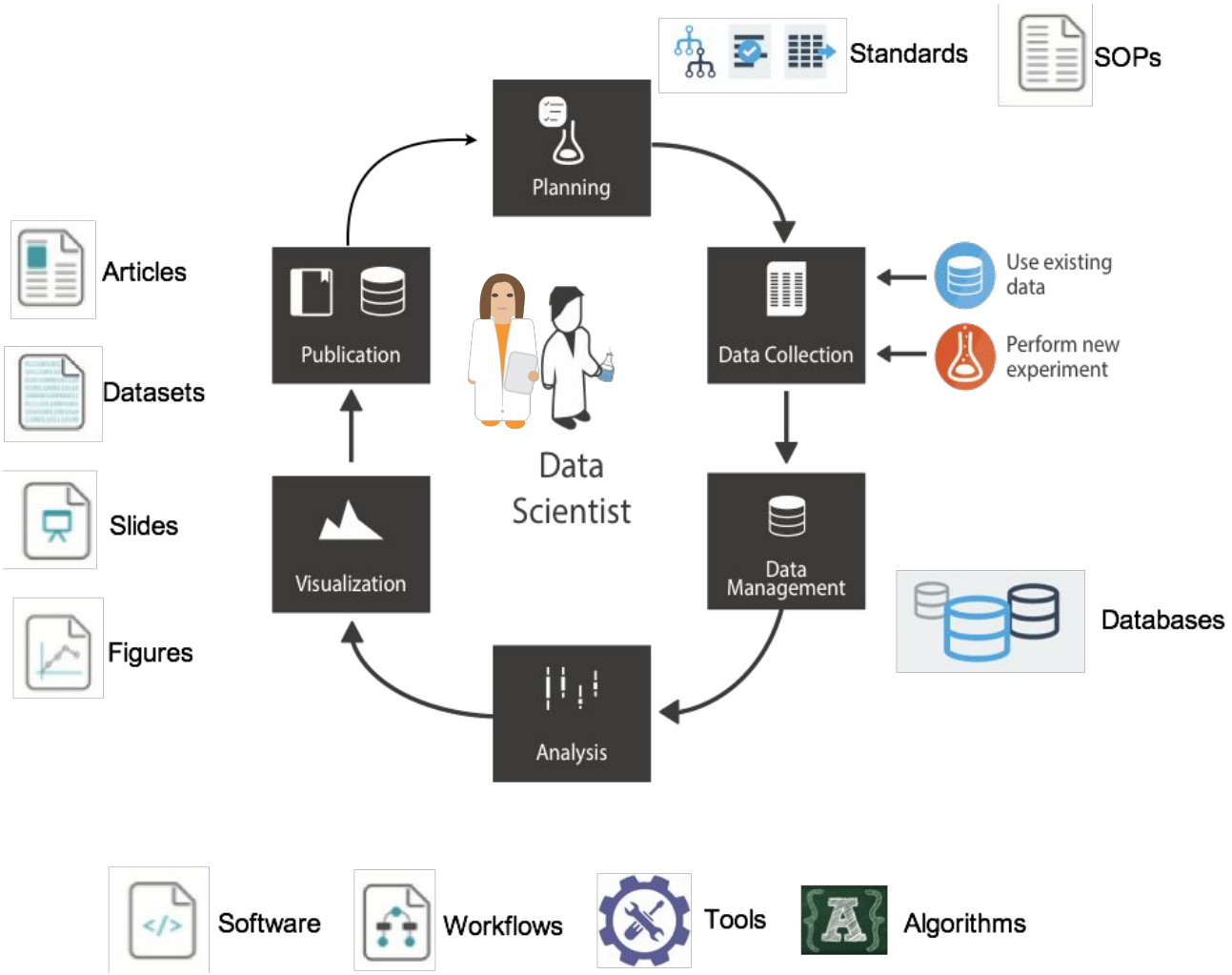
impa MR Lifia ATX Universidad Nacional de Rosario Queen's University Belfast UCL University of Oxford DATA CARPENTRY software carpentry Software Sustainability Institute W3C ESWC IS Auckland 2019 ISWC Science & Technology Facilities Council UK Research and Innovation

Research life cycle



consumer and producer of digital objects


Articles
Planning
Publication
Data Collection
Use existing data
Perform new experiment
Data Scientist
Visualization
Data Management
Analysis


Standards
SOPs
Planning
Articles
Datasets
Slides
Figures
Publication
Use existing data
Perform new experiment
Data Collection
Data Scientist
Visualization
Data Management
Databases
Analysis
Software
Workflows
Tools
Algorithms

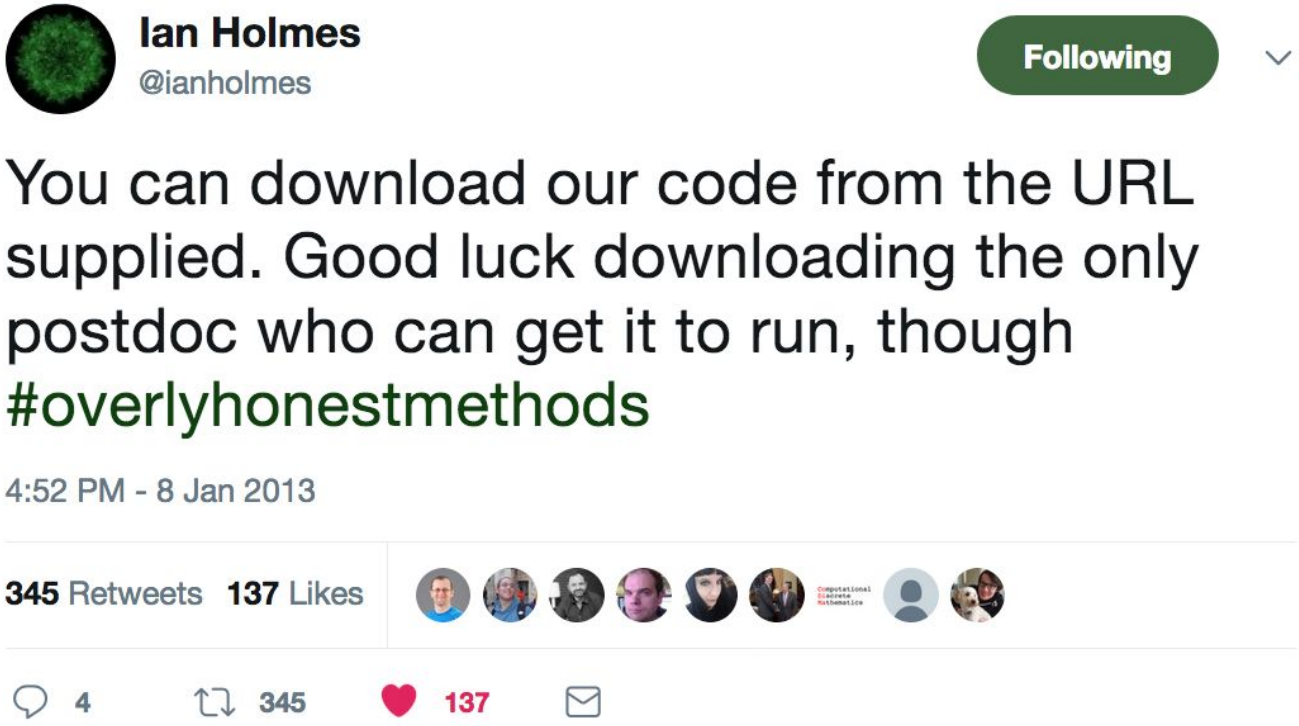
Better Software, Better Research

Research software

“Software that is **used to generate, process or analyse results that you intend to appear in a publication** (either in a journal, conference paper, monograph, book or thesis). Research software can be anything from a few lines of code written by yourself, to a professionally developed software package.”

“Software that does not generate, process or analyse results - such as word processing software, or the use of a web search - does not count as ‘research software’ [...].”

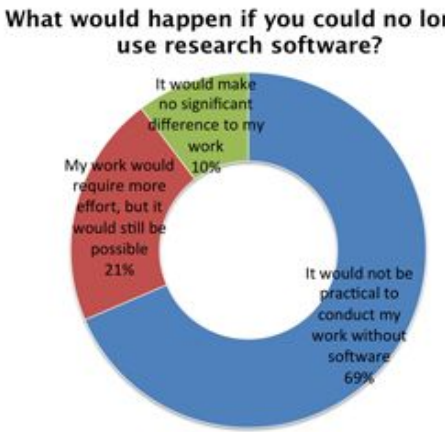
https://www.software.ac.uk/blog/2014-12-04-its-impossible-conduct-research-without-software-say-7-out-10-uk-researchers

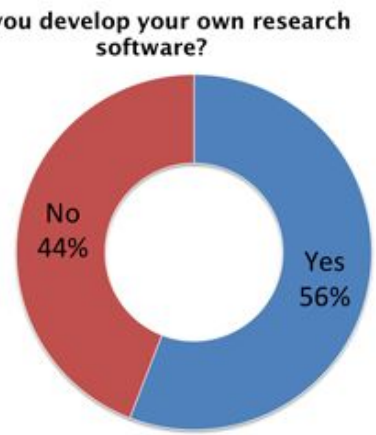


https://twitter.com/ianholmes/status/288689712636493824?lang=en

https://www.software.ac.uk/blog/2014-12-04-its-impossible-conduct-research-without-software-say-7-out-10-uk-researchers

No software, no research





417 researchers selected at random from 15 Russell Group universities

"S.J. Hettrick et al, UK Research Software Survey 2014", DOI:10.5281/zenodo.1183562

Software Sustainability Institute

http://software.ac.uk/

A national facility for cultivating better, more sustainable, research software to enable world-class research

- Software reaches boundaries in its development cycle that prevent improvement, growth and adoption
- Providing the expertise and services needed to negotiate to the next stage
- Developing the policy, forums and tools to support the research communities developing and using research software

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

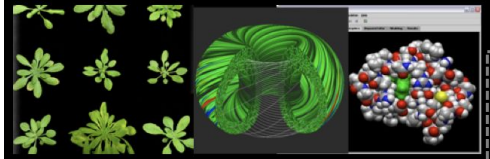
BBSRC

EPSRC Pioneering research and skills

Supported by EPSRC Grant EP/H043160/1
+ EPSRC/ESRC/BBSRC grant EP/N006410/1

Software

Helping the community to develop software that meets the needs of reliable, reproducible, and reusable research



Training

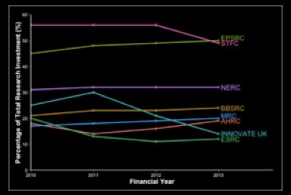
Delivering essential software skills to researchers via CDTs, institutions & doctoral schools

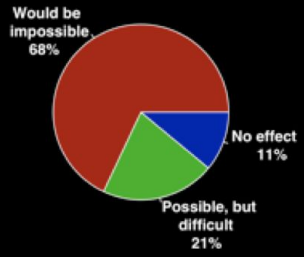
Outreach

Exploiting our platform to enable engagement, delivery & uptake

Policy

Collecting evidence on the community's software use & sharing with stakeholders

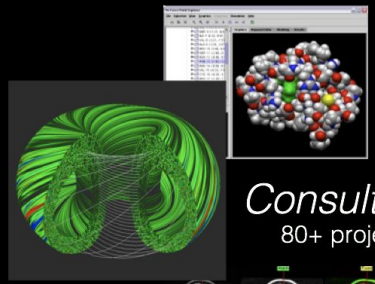




Community

Bringing together the right people to understand and address topical issues

Software



Consultancy

80+ projects

Advice

150+ evaluations
4 surgeries

Training

Courses

40+ UK SWC workshops
2000+ learners

Guides

90+ guides
50,000 readers

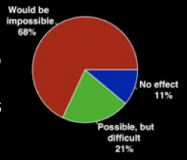
Outreach

Website & blog

200+ contributed articles
20,000 unique visitors per month
4,500 Twitter followers

Research

740 researchers
50,000 grants
analysed



Campaigns

BETTER SOFTWARE BETTER RESEARCH

1000+ RSEs engaged 2100 signatures

13 issues highlighted

Workshops

Journal of open research software

20+ workshops organised

Fellowship

112 domain ambassadors

Policy

Community

https://www.software.ac.uk/blog/2016-08-17-not-so-brief-history-research-software-engineers-0
https://rse.ac.uk/who/

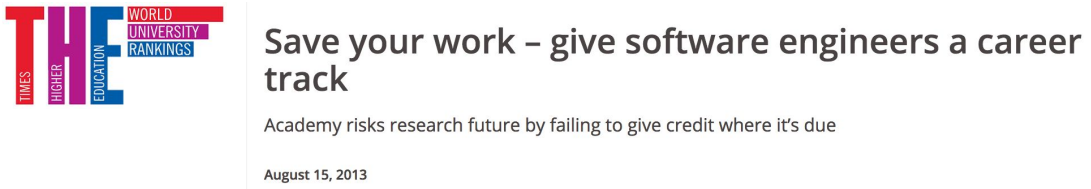
Research Software Engineer (RSE)

- People in research groups “who write code, not papers”
- March 2012 - SSI Collaborations Workshop - A new role in research
- Why is there no career for software developers in academia?
- Software is overlooked as merely “an uninteresting means of achieving interesting research”
- Software developers in academia lacked recognition but also lacked a name!
- Research Software Engineer
  - an understanding of both research and software engineering
- End 2013 - UK Research Software Engineers Association

RSE
RESEARCH SOFTWARE ENGINEERS ASSOCIATION

Research Software Engineers Association

https://rse.ac.uk



Save your work – give software engineers a career track

Academy risks research future by failing to give credit where it's due

August 15, 2013

https://www.timeshighereducation.com/news/save-your-work-give-software-engineers-a-career-track/2006431.article

Research Software Engineers Association

https://rse.ac.uk

Save your work – give software engineers a career track

Academy risks research future by failing to give credit where it's due

August 15, 2013

https://www.timeshighereducation.com/news/save-your-work-give-software-engineers-a-career-track/2006431.article

Recognising the Importance of Software in Research – Research Software Engineers (RSEs), a UK Example

Open Science Monitor Case Study

https://ec.europa.eu/info/files/recognising-importance-software-research-research-software-engineers-rses-uk-example_en

Research Software Engineers Association

https://rse.ac.uk

Save your work – give software engineers a career track

Academy risks research future by failing to give credit where it's due

August 15, 2013

https://www.timeshighereducation.com/news/save-your-work-give-software-engineers-a-career-track/2006431.article

Recognising the Importance of Software in Research – Research Software Engineers (RSEs), a UK Example

Open Science Monitor Case Study

https://ec.europa.eu/info/files/recognising-importance-software-research-research-software-engineers-rses-uk-example_en

Society of Research Software Engineering
@ResearchSoftEng

We are now the Society of Research Software Engineering! The Society has just been registered by the Charity Commission as an independent organisation to work for the improvements we all want to see. You will soon be able to join and shape the direction.

♡ 127 6:30 AM - Mar 14, 2019

https://twitter.com/ResearchSoftEng/status/1106125534105387009

ANNOUNCING THE CREATION OF THE SOCIETY OF RESEARCH SOFTWARE ENGINEERING

We are delighted to announce that the Society of Research Software Engineering is now a registered charity.

We applied last summer to register the new society with the Charity Commission as a Charitable Incorporated Organisation and we have just been told of their decision to approve this. Our Registered Charity Number is 1182455 and details will appear on their website shortly.

The UK RSE committee are now the trustees of the new Society and will be working to set up the organisation to meet our charitable aims.

This means that the RSE movement now has an independent organisation where the votes and input of members will shape the future direction. This is your society and we hope you'll play a part in determining how we, as a community, go forward to bring about the improvements we know are possible.

This is an important moment for everyone who believes in the value of Research Software Engineering. Reaching this point has been a true collective endeavour and would not have happened without the efforts of many people - from the earliest discussions at the start of the movement, through working to gain recognition for the role and build a community, to the recent push to drive through this application and establish a formal organisation.

Thank you to everyone who has registered their interest in joining the Society. We will be in touch as soon as we are ready to start accepting members.

The trustees of the Society

TANIA ALLARD
IAIN BETHUNE
ALYS BRETT
MIHAELA DUTA
ROB HAINES
SIMON HETTRICK
MATTHEW JOHNSON
ILIAN TODOROV
ANDY TURNER
MARK TURNER
CHRISTOPHER WOODS

SOCIETY OF RESEARCH SOFTWARE ENGINEERING

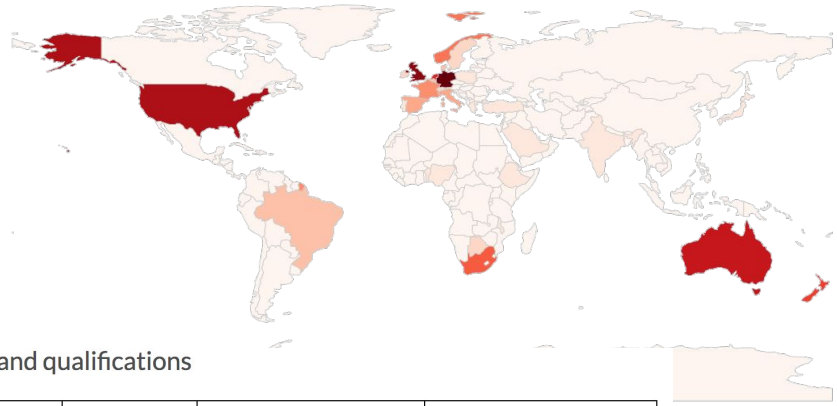
RSE Survey 2018

https://www.software.ac.uk/blog/2019-02-15-rse-survey-2018-results?mc_cid=25715ea091&mc_eid=b78c66ba86

https://www.software.ac.uk/what-do-we-know-about-rses-results-our-international-surveys

DOI 10.5281/zenodo.1194668

https://github.com/softwaresaved/international-survey



Discipline of study and work

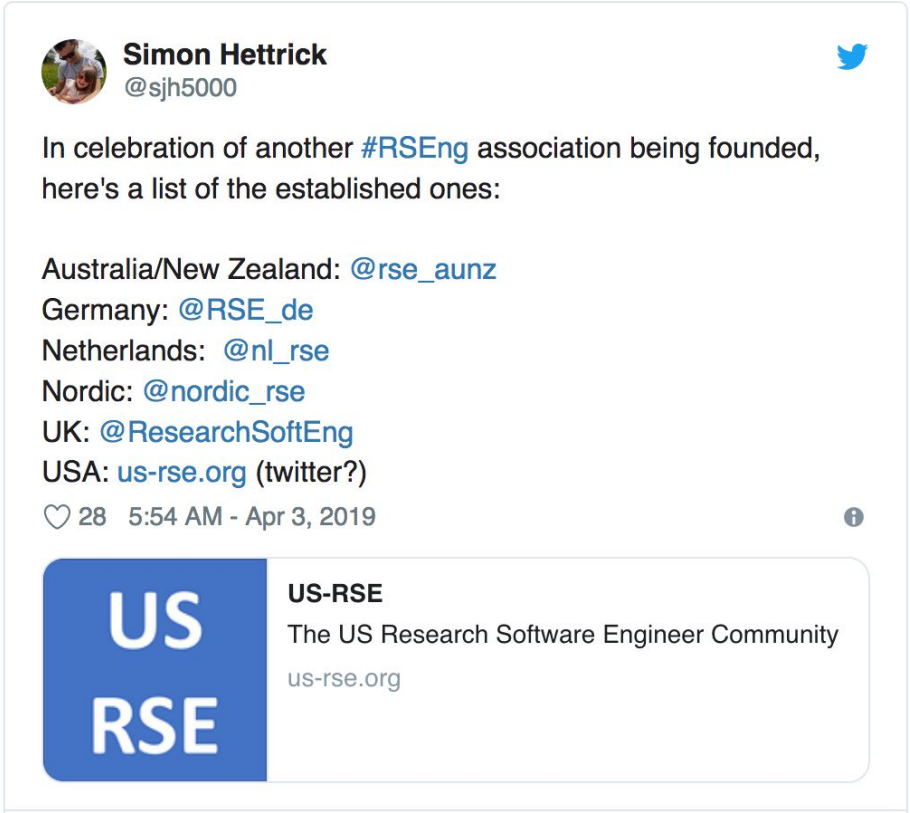
Countries	First discipline	Second discipline	Third discipline
Australia	Physical Sciences (28%)	Information and Computing Sciences (22%)	Biological Sciences (15%)
Germany	Computer science (26%)	Physics and astronomy (23%)	Geography & Environmental Sciences (8%)
Netherlands	Physics and Astronomy (33%)	Computer science (27%)	Biological Sciences (15%)
New Zealand	Computer Science (30%)	Biological Sciences (19%)	Physics and astronomy (11%)
South Africa	Computer Science (45%)	Biological Sciences (18%)	Mathematics (14%)
United Kingdom	Physics and Astronomy (34%)	Computer Science (24%)	Biological Sciences (12%)
United States	Physics and Astronomy (30%)	Computer Science (26%)	Biological Sciences (10%)
Rest of the World	Biological Sciences (18%)	Physics and astronomy (18%)	Computer Science (17%)

Gender, age and qualifications

Countries	Gender	Age	Qualification
Australia	Male (82%)	35 to 44 years (38%)	Doctorate (64%)
Germany	Male (86%)	25 – 44 (50%)	Master degree (51%)
Netherlands	Male (87%)	35 to 44 years (48%)	Doctorate (50%)
New Zealand	Male (91%)	35 to 44 years (29%)	Doctorate (37%)
South Africa	Male (89%)	35 to 44 years (47%)	Doctorate (30%)
United Kingdom	Male (80%)	35 to 44 years (40%)	Doctorate (70%)
United States	Male (73%)	25 to 34 years (32%)	Doctorate (45%)
Rest of the World	Male (92%)	25 to 34 years (19%)	Doctorate (64%)

RSEs Associations around the world

https://twitter.com/sjh5000/status/1113364116775866370



Version 1. F1000Res. 2017; 6: ELIXIR-876.
PMCID: PMC5490478
Published online 2017 Jun 13. doi: 10.12688/f1000research.11407.1

Four simple recommendations to encourage best practices in research software

Rafael C. Jiménez,[a,1] Mateusz Kuzak,[b,2] Monther Alhamdoosh,[3] Michelle Barker,[4] Bérénice Batut,[5] Mikael Borg,[6] Salvador Capella-Gutierrez,[7] Neil Chue Hong,[8] Martin Cook,[1] Manuel Corpas,[9] Madison Flannery,[10] Leyla Garcia,[11] Josep Ll. Gelpí,[12,13] Simon Gladman,[10] Carole Goble,[14] Montserrat González Ferreiro,[11] Alejandra Gonzalez-Beltran,[15] Philippa C. Griffin,[10] Björn Grüning,[5] Jonas Hagberg,[6] Petr Holub,[16] Rob Hooft,[17] Jon Ison,[18] Daniel S. Katz,[19,20,21,22] Brane Leskošek,[23] Federico López Gómez,[1] Luis J. Oliveira,[24] David Mellor,[25] Rowland Mosbergen,[26] Nicola Mulder,[27] Yasset Perez-Riverol,[11] Robert Pergl,[28] Horst Pichler,[29] Bernard Pope,[10] Ferran Sanz,[30] Maria V. Schneider,[10] Victoria Stodden,[20] Radosław Suchecki,[31] Radka Svobodová Vařeková,[32,33] Harry-Anton Talvik,[34] Ilian Todorov,[35] Andrew Treloar,[36] Sonika Tyagi,[10,37] Maarten van Gompel,[38] Daniel Vaughan,[11] Allegra Via,[39] Xiaochuan Wang,[40] Nathan S. Watson-Haigh,[31] and Steve Crouch[c,41]

1. Make source code publicly accessible from day one

2. Make software easy to discover by providing software metadata via a popular community registry

3. Adopt a licence and comply with the licence of third-party dependencies

4. Define clear and transparent contribution, governance and communication processes

Software Publication

The Journal of Open Source Software

Software X

Software Sustainability Institute

Journal of open research software

Concurrency and Computation Practice and Experience

Journal of: Software: Practice and Experience

Software Citation

nature
International journal of science

Toolbox

Toolbox | 23 April 2019

A simple approach to dating bones

Forensic anthropologist Ann Ross describes the techniques she uses to determine the age of human skeletons.

Jeffrey M. Perkel

Toolbox | 01 April 2019

11 ways to avert a data-storage disaster

Hard-drive failures are inevitable, but data loss doesn't have to be.

Jeffrey M. Perkel

RIO

https://www.software.ac.uk/which-journals-should-i-publish-my-software

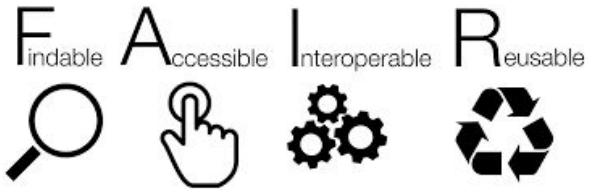
Better Data,
Better Research

SCIENTIFIC DATA

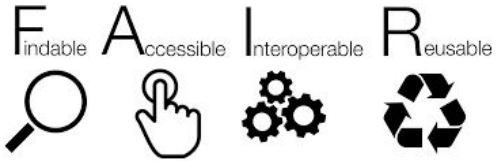
Comment | OPEN | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons ✉ - Show fewer authors



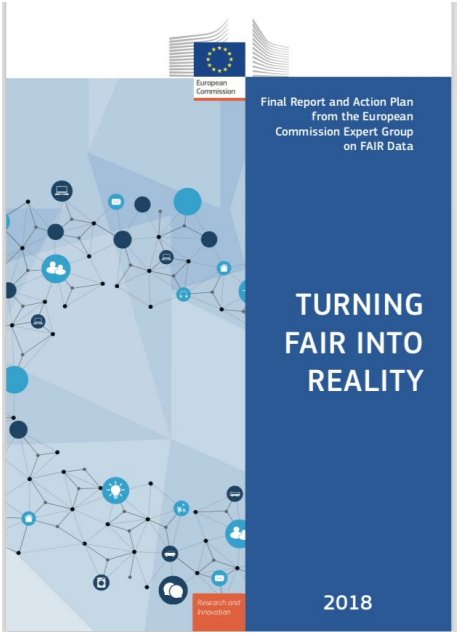
Scientific Data **3**, Article number: 160018 (2016) | Download Citation ⬇

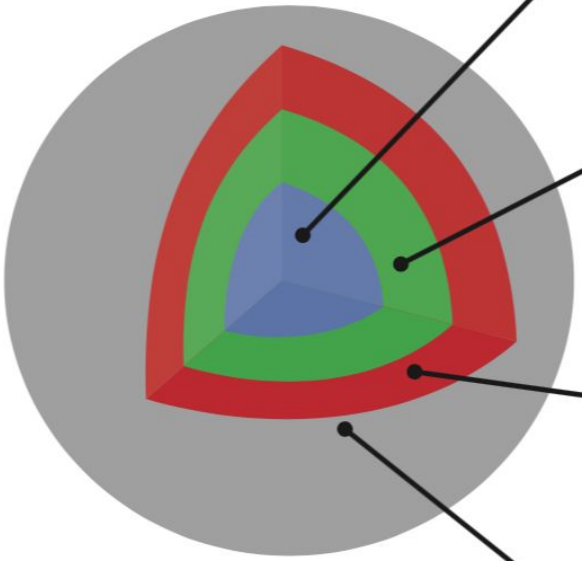


Data Principles

Emphasis is on enhancing the ability of ***machines*** to automatically find and use the data, in addition to supporting its reuse by ***individual***

- **Findable**
  - Globally unique, resolvable, and persistent identifiers
  - Machine-readable metadata to support structured search
- **Accessible**
  - Clearly defined access and security protocols
- **Interoperable**
  - Extensible machine interpretable formats for data + metadata
  - Linked to other resources
- **Reusable**
  - Provide licensing, provenance, and use community-standards





DIGITAL OBJECT

Data, code and other research outputs

At its most basic level, data or code is a bitstream or binary sequence. For this to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and documentation. These layers of meaning enrich the object and enable reuse.

IDENTIFIERS

Persistent and unique (PIDs)

Digital Objects should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE

Open, documented formats

Digital Objects should be represented in common and ideally open file formats. This enables others to reuse them as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.

METADATA

Contextual documentation

In order for Digital Objects to be assessable and reusable, they should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the objects were created. To enable the broadest reuse, they should be accompanied by a plurality of relevant attributes and a clear and accessible usage license.

https://biocaddie.org/

bioCADDIE

Build a PubMed for Data

dataMED BETA version

JATS (Journal Article Tag Suite) underpins **PubMed** for literature indexing,

DATS (DatA Tag Suite) the model to index data sources

(used by **DataMed**, **but not limited to it**)

DATS
DatA Tag Suite

Funded by the United States National Institutes of Health (NIH)

dataMED BETA version

biomedical and healthCAre Data Discovery Index Ecosystem bioCADDIE

About | Feedback | Web API | Submit | Login

Engaging The Community Toward a Data Discovery Index (DataMed v3.0)

Search for data through bioCADDIE

Search for data set | Search for repository

Advanced Search help

Statistics

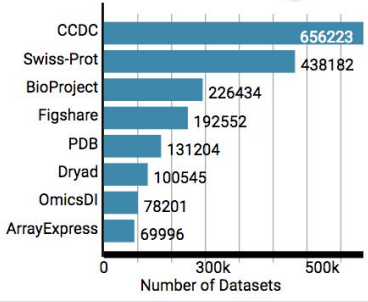
75 REPOSITORIES

15 DATA TYPES

2,336,403 DATASETS

4 PILOT PROJECTS

Top 8 Repositories



Top 7 Datasets

116 The effect of growth rate on pyrazinamide activity in Mycobacterium tuberculosis; insights for early bactericidal activity

114 Effect of estrogen receptor- or retinoic acid receptor-knockdown on estrogen-sensitive MCF7 breast cancer cells

95 NM_000229.1(LCAT):c.440C>T (p.Thr147Ile) AND Fish-eye disease

82 Block design food and nonfood picture viewing task

78 Coronaviridae : Coronaviridae

74 Sleep Heart Health Study

72 Time Magazine/Abt SRBI Poll # 2008-4556: Race

New Features

July 31, 2017. v3.0

» API for DataMed
» Autocompleting of words
» Reporting of broken links
» Context for search terms
» Increase in number of repositories that are indexed

Feb 28, 2017. v2.0

» Increase coverage to more repositories
» Duplicate datasets display feature
» Usability enhancements based on user feedback and user interviews
» User-reported issues resolved

https://datamed.org/

https://fairsharing.github.io/JSONschema-documenter/?schema_url=https://w3id.org/dats/schema/dataset_schema.json#

DATS Dataset Schema

Schema Metadata

- **$schema :** http://json-schema.org/draft-04/schema
- **id :** https://w3id.org/dats/schema/dataset_schema.json
- **title :** DATS Dataset Schema
- **description :** A set of dimensions about an entity being observed. A collection of data, published or curated by a single agent, and available for access or download in one or more formats (from DCAT: http://www.w3.org/TR/vocab-dcat/#Class:_Dataset). A body of structured information describing some topic(s) of interest (from: http://schema.org/Dataset).
- **type :** object
- **additionalProperties :** false
- **required :** ["title","types","creators"]

Schema Fields

@context

Description: The JSON-LD context

Expected types (any number of types from below):

- **string**
- **object**
- array

@id

Description: The JSON-LD identifier

Expected type : uri

@type

Description: The JSON-LD type

Expected value(s) :

- Dataset

Expected type : string

identifier

Cardinality: (0 ... 1)

Expected type:

identifier_info_schema

alternateIdentifiers

Description: Alternate identifiers for the dataset.

Cardinality: (0 ... n)

relatedIdentifiers

Description: Related identifiers for the dataset.

Cardinality: (0 ... n)

nature genetics

http://dx.doi.org/10.1038/ng.3864

Finding useful data across multiple biomedical data repositories using DataMed

Lucila Oh
Hua Xu[7,8]
Ergin Soy

The value
research c
with an int
index and
DataMed i
findability
compose t

www.nature.com/scientificdata

http://doi.org/10.1038/sdata.2017.59

SCIENTIFIC DATA

OPEN

DATS, the data tag suite to enable discoverability of datasets

Susanna-Assunta Sanson
George Alter[2], Jeffrey S.
Xiaoling Chen[4], Hyeon-e
Lucila Ohno-Machado[3]

Received: 27 January 2017
Accepted: 30 March 2017
Published: 6 June 2017

Today's science increasingly
distributed across a range of
may soon become as essen
international collaborative
Knowledge (BD2K) initiative
support the DataMed data d
the scientific literature. Akir
enables submission of meta
generic and applicable to a
specialized data types. DAT
serialization in schema.org,
Microsoft, Yahoo and Yand

Journal of the American Medical Informatics Association, 25(1), 2018, 13–16
doi: 10.1093/jamia/ocx119
Advance Access Publication Date: 8 December 2017
Brief Communication

AMIA OXFORD

https://doi.org/10.1093/jamia/ocx119

Brief Communication

Data discovery with DATS: exemplar adoptions and lessons learned

Alejandra N Gonzalez-Beltran,[1,*] John Campbell,[2] Patrick Dunn,[2] Diana Guijarro,[3] Sanda Ionescu,[4] Hyeoneui Kim,[3] Jared Lyle,[4] Jeffrey Wiser,[2] Susanna-Assunta Sansone,[1] and Philippe Rocca-Serra[1,*]

W3C Editor's Draft

TABLE OF CONTENTS

1. **Introduction**
2. **Motivation for change**
3. **Namespaces**
4. **Conformance**
5. **Vocabulary overview**
   - 5.1 DCAT scope
   - 5.2 RDF considerations
   - 5.3 Basic Example
   - 5.4 Classifying datasets thematically
   - 5.5 Classifying dataset types
   - 5.6 Describing catalog records metadata
   - 5.7 Dataset available only behind some Web page
   - 5.8 A dataset available as a download and behind some Web page
   - 5.9 A dataset available through a service
6. **Vocabulary specification**
   - 6.1 RDF representation
   - 6.2 Elements from other vocabularies
     - 6.2.1 Complementary vocabularies
     - 6.2.2 Element definitions
   - 6.3 Class: Catalog
     - 6.3.1 Property: homepage

Data Catalog Vocabulary (DCAT) - Revised edition

ReSpec

W3C Editor's Draft 20 April 2019

This version:
https://w3c.github.io/dxwg/dcat/

Latest published version:
https://www.w3.org/TR/vocab-dcat-2/

Latest editor's draft:
https://w3c.github.io/dxwg/dcat/

Latest Recommendation:
https://www.w3.org/TR/vocab-dcat/

Editors:
Riccardo Albertoni (CNR - Consiglio Nazionale delle Ricerche, Italy)
Dave Browning (Refinitiv)
Simon Cox (CSIRO)
Alejandra Gonzalez Beltran (Oxford eResearch Centre, Engineering Science, University of Oxford)
Andrea Perego (European Commission, Joint Research Centre)
Peter Winstanley (Scottish Government)

Participate:
GitHub w3c/dxwg
File a bug
Commit history
Pull requests

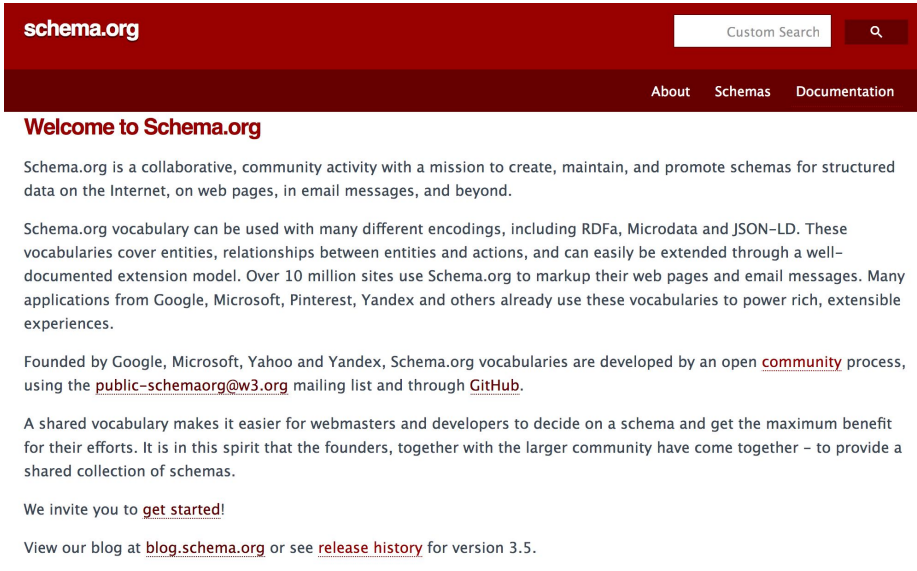
Contributors:
Makx Dekkers

Google Dataset Search Beta

Try boston education data or weather site:noaa.gov

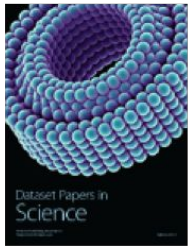
Learn more about including your datasets in Dataset Search.

https://toolbox.google.com/datasetsearch



Data Publication

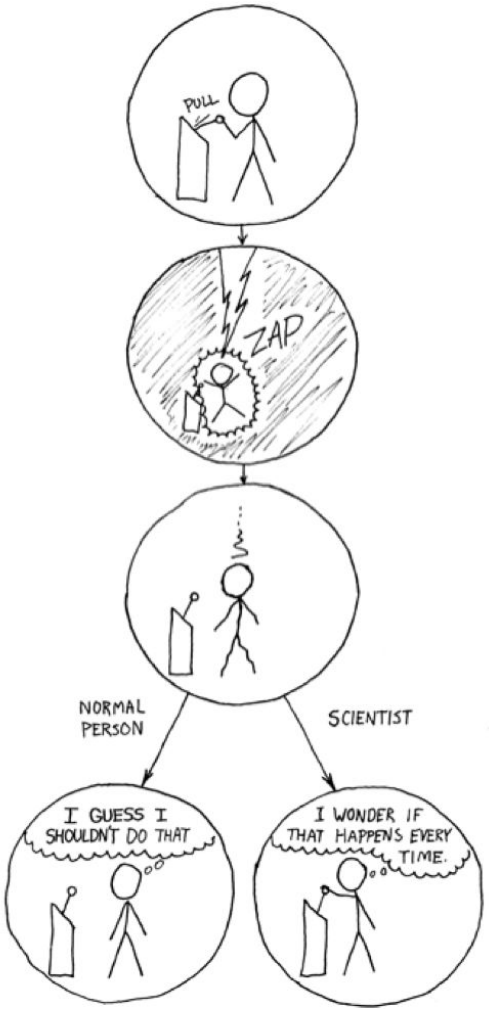
Data Citation



- Incentive, credit for sharing
  - Big and small data
  - Unpublished data
  - Long tail of data
  - Curated aggregation
- Peer review of data
- Value of data vs. analysis
- Discoverability and reusability
  - Complementing community databases

Better Software +
Better Data =
Better Research

Research reproducibility



http://xkcd.com/242

A few definitions...

Research results can be considered **established** when they can be **repeated**, **replicated** and ultimately independently **reproduced**

Plesser, H. E. (2017). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, *11*, 76. http://doi.org/10.3389/fninf.2017.00076

A few definitions...

Research results can be considered **established** when they can be **repeated**, **replicated** and ultimately independently **reproduced**

–**Repeatability** (same team, same experimental setup)

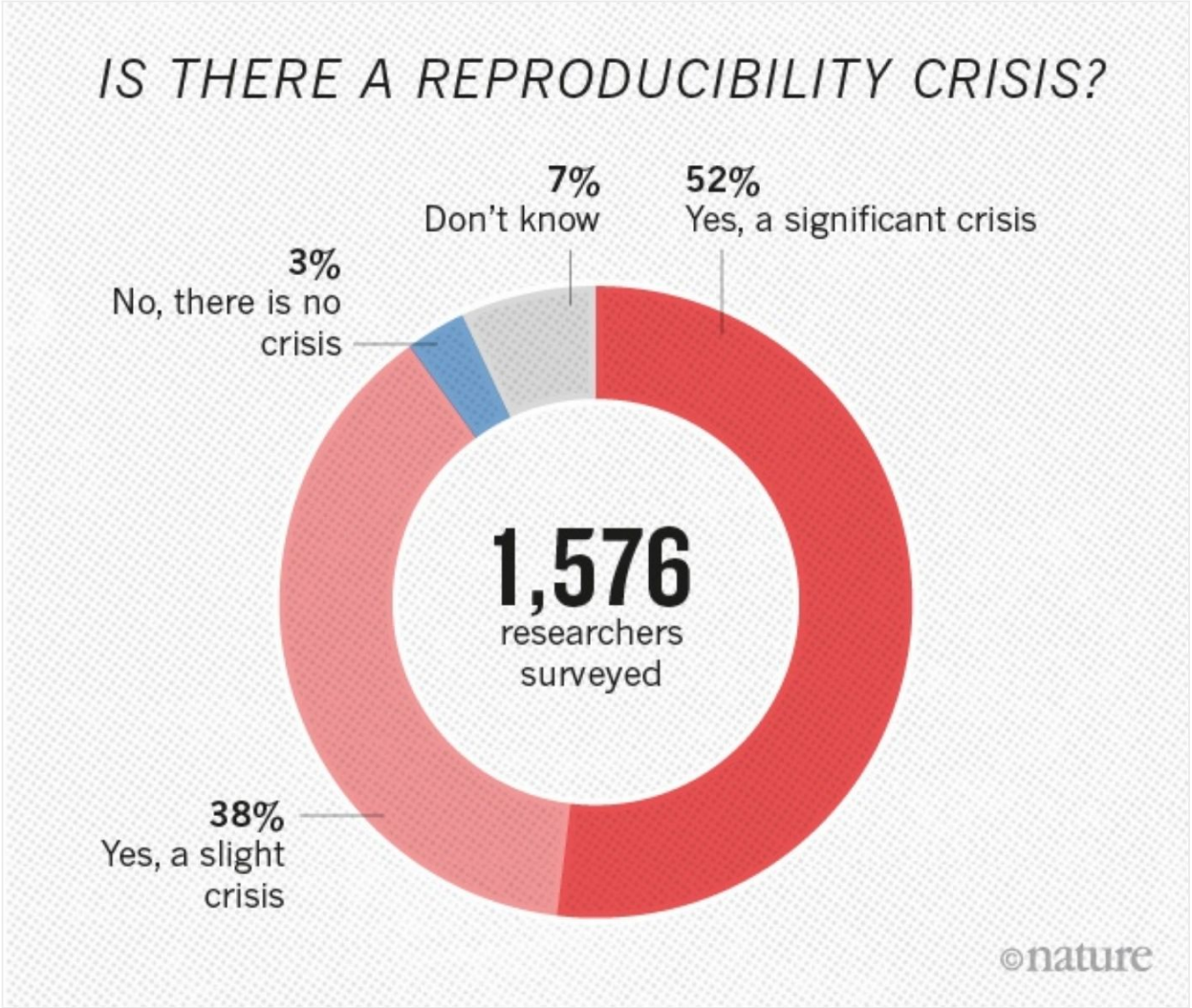
–**Replicability** (different team, same experimental setup)

–**Reproducibility** (different team, different experimental setup)

Plesser, H. E. (2017). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, *11*, 76. http://doi.org/10.3389/fninf.2017.00076

The reproducibility crisis

http://petcaretips.net/bonding-rabbit-to-pets.html



NATURE | VOL 533 | 26 MAY 2016

NATURE | NEWS FEATURE

1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

Monya Baker

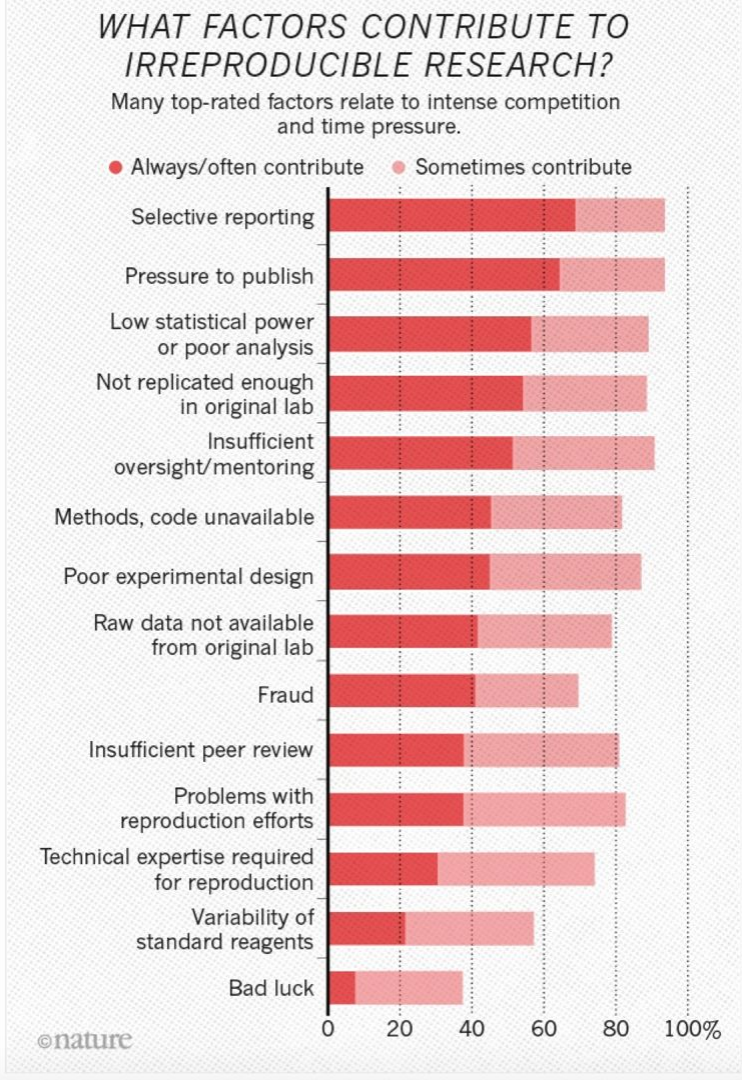
25 May 2016

Why?

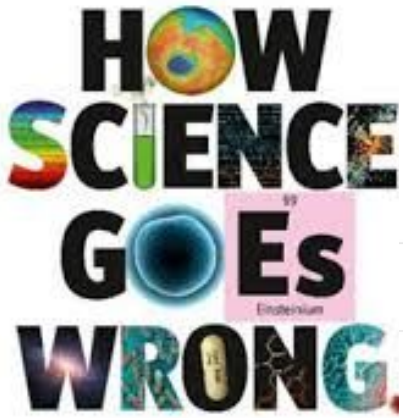
- Data growth
- Digitalization of scholarly resources
- Technological advances
- Data-driven, quantitative research
- Lack of data management and computational skills
- Lack of incentives

Why?

- Data growth
- Digitalization of scholarly resources
- Technological advances
- Data-driven, quantitative research
- Lack of data management and computational skills
- Lack of incentives



The reproducibility crisis goes mainstream



Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

A SIMPLE idea underpins science: "trust, but v
always be subject to challenge from experime
powerful idea has generated a vast body of kn

Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

Scientific publishing

Are research papers less accurate and truthful than in the past?

That is a myth, according to the latest investigation

https://www.economist.com/science-and-technology/2018/03/17/are-research-papers-less-accurate-and-truthful-than-in-the-past

“Mistakes are part of normal science”

Commentary:
Fallibility in science: Responding to errors in the work of oneself and others

Bishop, Dorothy V. **PeerJ PrePrints; San Diego** (Dec 24, 2017).
DOI:http://doi.org10.7287/peerj.preprints.3486v1

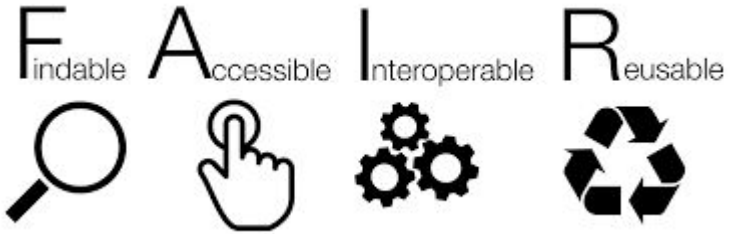
- Dilemmas for scientists who find errors in their own work
- Dilemmas for scientists who find errors in the work of others
- Perceptions and realities of reputational damage
- Best defense: adoption of open science practices
- For science to progress, we have to accept the inevitability of errors

Two strands for **reproducibility**...

DATA

SOFTWARE

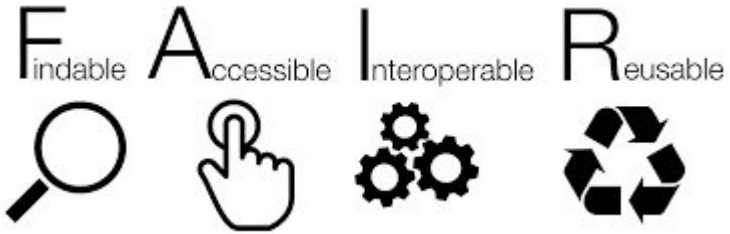
Two strands for **reproducibility**...



DATA

SOFTWARE

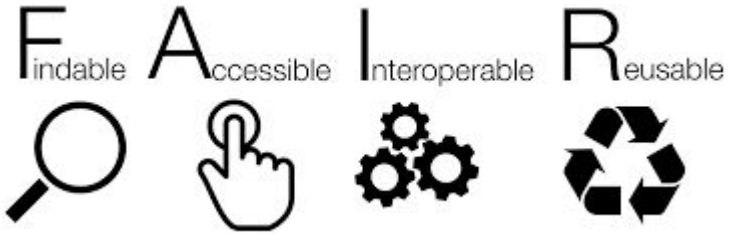
Two strands for **reproducibility**...



DATA

SOFTWARE

Two strands for **reproducibility**...



DATA

SOFTWARE

Experiments in reproducibility

Reproducibility Challenge within the International Semantic Web Conference (ISWC)

Alejandra Gonzalez-Beltran & Michael Cochez

- The **International Semantic Web Conference** (https://iswc2019.semanticweb.org/) is the premier international forum for the Semantic Web and Linked Data community; 18th edition happening in New Zealand in October this year
- Topics for this year: knowledge graphs, linked data, linked schemas and AI on the Web
- For the first time, we are running a **Reproducibility Challenge** for ISWC submissions
  - Authors of accepted research track papers that include a significant experimental evaluation will be invited to participate in a reproducibility certification (with multiple levels)
  - Dedicated Programme Committee of '**reproducers'**
  - **'Reproducers'** are encouraged to interact with the authors to evaluate their submissions
  - The SIGMOD conference has been running a similar challenge
  - The evaluation implies trying to replicate and/or reproduce the submission's results with the help of the authors
  - Objectives
    - Encourage a culture of data and code sharing
    - Encourage a culture of rigorous and transparent results
    - Highlight the impact of semantic web research
    - Enable easy dissemination of results
  - We are considering **different levels of reproducibility** and we are **looking for feedback -** possible levels are:
    - Data available (e.g. via Data Use Agreement) & software available
    - Replicated results
    - Reproducible results

What levels should we consider? How we differentiate between them?

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

From Peer-Reviewed to Peer-Reproduced in Scholarly Publishing: The Complementary Roles of Data Models and Workflows in Bioinformatics

Alejandra González-Beltrán, Peter Li, Jun Zhao, Maria Susana Avila-Garcia, Marco Roos, Mark Thompson, Eelke van der Horst, Rajaram Kaliyaperumal, Ruibang Luo, Tin-Lap Lee, Tak-wah Lam, Scott C. Edmunds, Susanna-Assunta Sansone, Philippe Rocca-Serra

Published: July 8, 2015 • DOI: 10.1371/journal.pone.0127612

Abstract
Introduction
Results
Discussion
Systems and Methods
Conclusion

Abstract

Motivation

Reproducing the results from a scientific paper can be challenging due to the absence of data and the computational tools required for their analysis. In addition, details relating to the procedures used to obtain the published results can be difficult to discern due to the use of natural language when reporting how experiments have been performed. The

Can **data models** and **computational workflows** help in capturing the experimental processes and **reproduce findings**? How?

Technical Note

Highly accessed | Open Access

SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler

Ruibang Luo[1,2†], Binghang Liu[1,2†], Yinlong Xie[1,2,3†], Zhenyu Li[1,2], Weihua Huang[1], Jianying Yuan[1], Guangzhu He[1], Yanxiang Chen[1], Qi Pan[1], Yunjie Liu[1], Jingbo Tang[1], Gengxiong Wu[1], Hao Zhang[1], Yujian Shi[1], Yong Liu[1], Chang Yu[1], Bo Wang[1], Yao Lu[1], Changlei Han[1], David W Cheung[2], Siu-Ming Yiu[2], Shaoliang Peng[4], Zhu Xiaoqian[4], Guangming Liu[4], Xiangke Liao[4], Yingrui Li[1,2], Huanming Yang[1], Jian Wang[1], Tak-Wah Lam[2*] and Jun Wang[1*]

(GIGA)n SCIENCE

Evaluation of SOAPdenovo2 tool for the de novo assembly of genomes from small DNA segments reads by next generation sequencing, implementing improvements over SOAPdenovo1 assembler.

- open peer-review
- availability of data
- analysis scripts
- documentation

pre-publication history

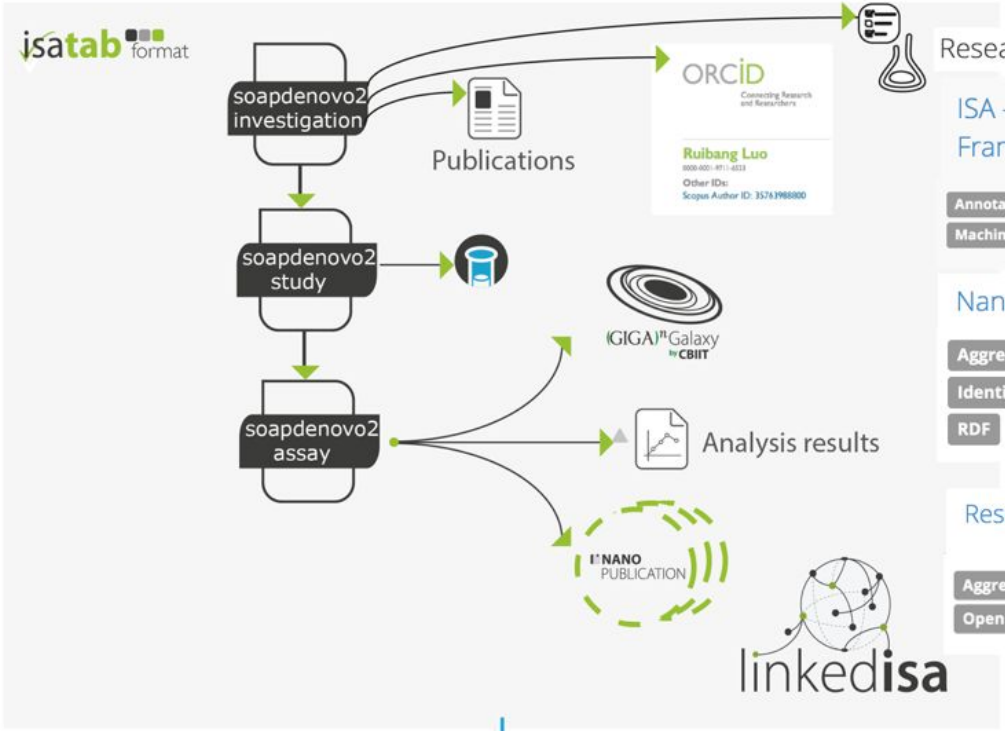
Original Submission - Version 1
Reviewer's Report
Reviewer's Report
Reviewer's Report
Resubmission - Version 2
Reviewer's Report
Reviewer's Report

http://sourceforge.net/projects/soapdenovo2

https://github.com/aquaskyline/SOAPdenovo2

github SOCIAL CODING

Aggregation and workflow preservation as researchobject



Research Object Initiatives & Resources

ISA – Investigation, Study, Assay Framework

Annotation | Commons | Description | Interoperable | ISA | Machine Readable | RDF

Nanopublications

Aggregation | Annotation | Description | Identity | Machine Readable | Provenance | RDF

Research Object (RO) Model

Aggregation | Annotation | Identity | JSON | OAI-ORE | Open Annotation | RDF

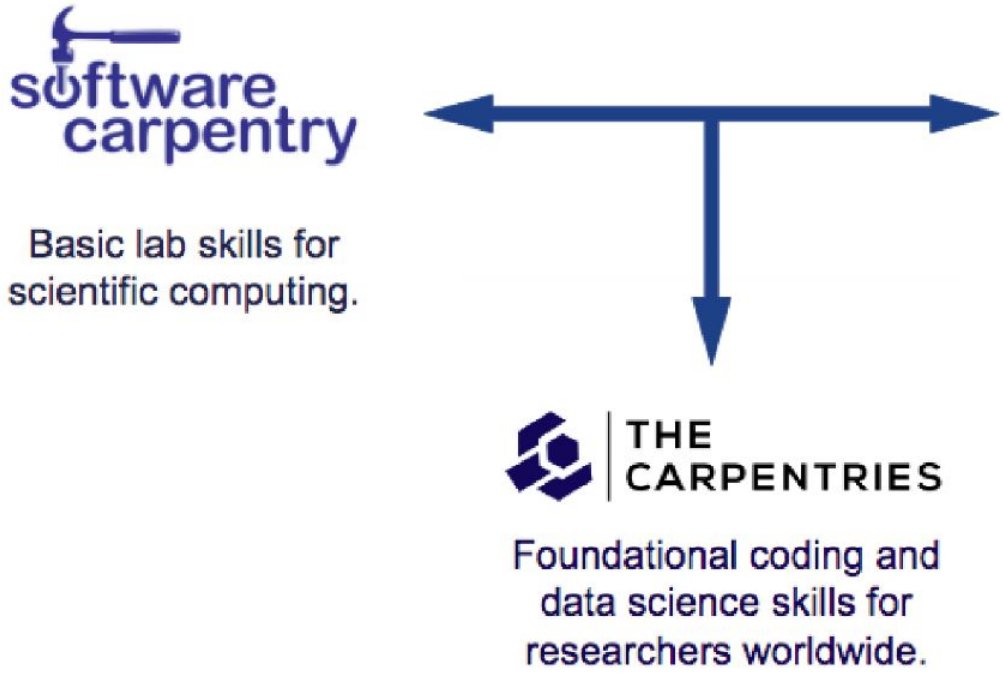
http://www.researchobject.org/

http://isa-tools.github.io/soapdenovo2

Main thoughts about the reproducibility experiment

- Use of Virtual Research Environments to re-enact and validate data analysis
  - Galaxy, Jupyter notebooks / Binder
  - Features: data manipulation, editing and document hosting
- Collaborative science: data and analytical workflows sharing
- Tools for software packaging, platform virtualisation
- Reporting of scientific work improved by the use of data models
  - Complementary models: ISA, nanopublications, research object
  - Enables sharing, reuse, repurposing
- Need for re-evaluation of tools for scholarly publishing
- Costs of peer-review, costs of reproducibility
- Reproducibility as a spectrum

Skills & Capacity Building
Training in the gaps



Teal, Tracy (2018): GCCBOSC 2018 Democratizing Data talk. figshare. Presentation.

https://doi.org/10.6084/m9.figshare.6709493.v1

- Code skills for effective research computing
- Trained instructors
- Two-day hands-on workshops
- Collaboratively developed, openly licensed lesson materials

THE CARPENTRIES Impact.

According to Programmatic & Long-Term Survey Respondents

- 1,480 Instructors badged
- 1,332 Workshops taught
- 37,000 Learners reached
- 44 countries
- ...and growing every year!

- 77% are more comfortable with tools
- 54% have made their analysis more reproducible
- 65% have gained confidence working with data
- 74% have recommended The Carpentries workshops to friends or colleagues.

Teal, Tracy (2018): GCCBOSC 2018 Democratizing Data talk. figshare. Presentation.

https://doi.org/10.6084/m9.figshare.6709493.v1

THE CARPENTRIES

Lessons

Curriculum

Our lessons are developed collaboratively on GitHub. You can check the status of each lesson on our dashboard, or look at older releases.

Availability

Contributing

If you have questions about contributing to our lessons, visit each lesson's GitHub repo to submit an issue or to get the link to join that lesson's Maintainers' discussion on Slack. For general information on how to contribute to our lessons, see our contributors guide. To learn more about how our lessons are structured, and why, please see the example lesson.

https://software-carpentry.org/lessons/

Our Core Lessons in English

Lesson	Site	Repository	Reference	Instructor Guide	Maintainer(s)
The Unix Shell					Gabriel Devenyi, Ashwin Srinath, Colin Morris, Will Pitchers
Version Control with Git					Ivan Gonzalez, Daisie Huang, Nima Hejazi, Katherine Koziar, Madicken Munk
Programming with Python					Trevor Bekolay, Valentina Staneva, Anne Fouilloux, Maxim Belkin, Mike Trizna
Plotting and Programming in Python					Nathan Moore, Allen Lee, Sourav Singh, Olav Vahtras
Programming with R					Daniel Chen, Katrin Leinweber, Diya Das
R for Reproducible Scientific Analysis					Thomas Wright, Naupaka Zimmerman, Jeffrey Oliver, David Mawdsley

software carpentry

https://software-carpentry.org/lessons/

Our Core Lessons in Spanish

Lesson	Site	Repository	Reference	Instructor Guide	Maintainer(s)
La Terminal de Unix					Ivan Gonzalez, Clara Llebot, Verónica Jiménez, Silvana Pereyra, Heladia Salgado
Control de versiones con Git					Ivan Gonzalez, Rayna Harris, Clara Llebot
R para Análisis Científicos Reproducibles					Rayna Harris, Verónica Jiménez, Silvana Pereyra, Heladia Salgado

Main take-home messages

- Research Lifecycle
  - Research outputs, sharing incentives, credit
- Better Software
  - Software sharing
  - Software Sustainability, Credit
  - Building communities of practice: SSI / RSEs / Carpentries
  - Recognition of research software and research software engineers
- Better Data
  - Data sharing
  - FAIR (Findable, Accessible, Interoperable and Reusable) data
- Better Research
  - Reproducibility is an issue across domains - reproducibility crisis
  - Reproducibility Challenge - ISWC