

# Data FAIRification using RStudio workflows

---

Brendan Palmer & Darren Dahly,  
Clinical Research Facility - Cork & School of Public Health,  
University College Cork  
 [@B\\_A\\_Palmer](https://twitter.com/B_A_Palmer) [@statsepi](https://twitter.com/statsepi)

# Where to begin...



# Don't do what Donny Dont does!



"In short, peer review misses all the hard stuff, and a worrying amount of the easy stuff"

James Heathers,  
Northwestern University

#datathugs



## Brian Wansink: The grad student who never said no

"Every day we would scratch our heads, ask "Why," and come up with another way to reanalyze the data with yet another set of plausible hypotheses. Eventually we started discovering solutions"

# Credibility crisis

2005

PLOS MEDICINE

BROWSE PUBLISH ABOUT SEARCH advanced search

OPEN ACCESS ESSAY

## Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • <https://doi.org/10.1371/journal.pmed.0020124>

68,436 Save	3,184 Citation
2,813,238 View	10,483 Share

2016

nature International weekly journal of science

Search Go Advanced search

Home News & Comment Research Careers & Jobs Current Issue Archive Audio & Video For Authors

Archive Volume 533 Issue 7604 News Feature Article

NATURE | NEWS FEATURE

1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

Monya Baker

2018

## THE IRREPRODUCIBILITY CRISIS OF MODERN SCIENCE

Causes, Consequences, and the Road to Reform



DAVID RANDALL AND CHRISTOPHER WELSER  
NATIONAL ASSOCIATION OF SCHOLARS  
APRIL 2018  
ISBN: 978-0-9986635-5-5



# REFLECTIONS

ON THE

## DECLINE OF SCIENCE IN ENGLAND,

AND ON

SOME OF ITS CAUSES.

BY

CHARLES BABBAGE, ESQ.

LUCASIAN PROFESSOR OF MATHEMATICS IN THE UNIVERSITY OF CAMBRIDGE,  
AND MEMBER OF SEVERAL ACADEMIES.

LONDON:

PRINTED FOR B. FELLOWES, LUDGATE STREET;  
AND J. BOOTH, DUKE STREET, PORTLAND PLACE.

1830

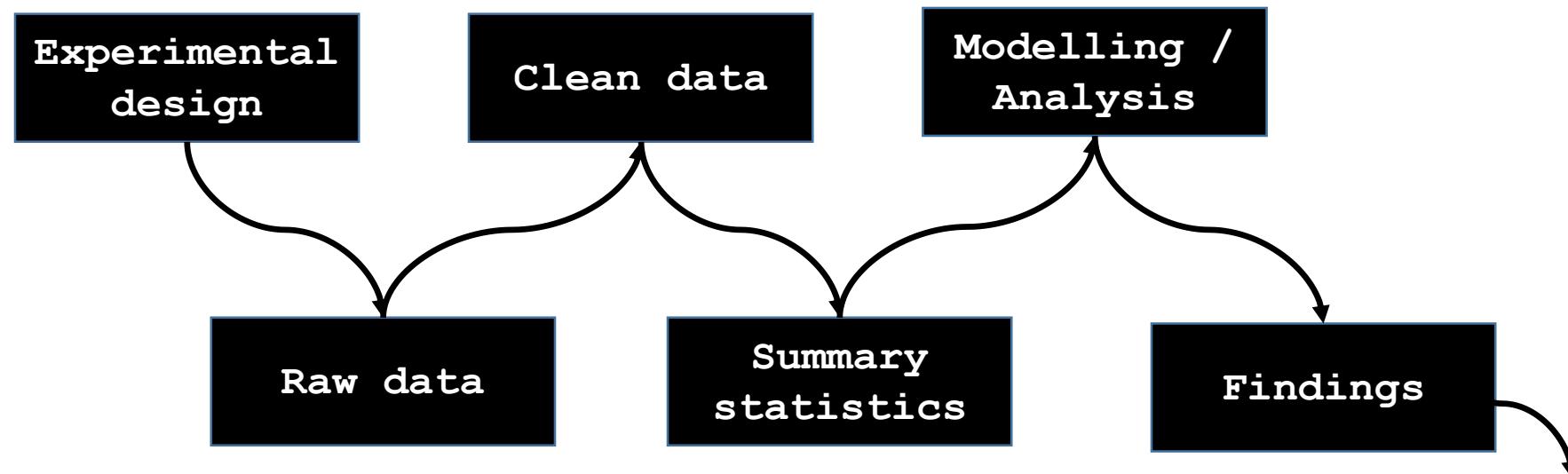
# p-values should not define a study

The screenshot shows the homepage of the journal **nature**, which is described as an "International weekly journal of science". The top navigation bar includes links for Home, News & Comment, Research, Careers & Jobs, Current Issue, Archive, Audio & Video, and For Authors. Below this, a breadcrumb navigation shows the path: Archive > Volume 520 > Issue 7549 > Comment > Article. On the right side of the header are search fields, a "Go" button, and links for Advanced search, E-alert, RSS, Facebook, and Twitter. The main content area features the title "Statistics: *P* values are just the tip of the iceberg" in large, bold letters, followed by the authors' names, Jeffrey T. Leek & Roger D. Peng, and the date, 28 April 2015.

## Statistics: *P* values are just the tip of the iceberg

Jeffrey T. Leek & Roger D. Peng

28 April 2015

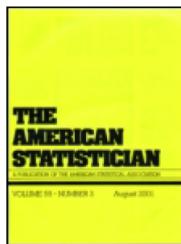


# The winds of change

## CONSORT 2010

The CONSORT (CONsolidated Standards of Reporting Trials) 2010 guideline is intended to improve the reporting of parallel-group randomized controlled trial (RCT), enabling readers to understand a trial's design, conduct, analysis and interpretation, and to assess the validity of its results. This can only be achieved through complete adherence and transparency by authors.

CONSORT 2010 was developed through collaboration and consensus between clinical trial methodologists, guideline developers, knowledge translation specialists, and journal editors (see [CONSORT group](#)). CONSORT 2010 is the current version of the guideline and supersedes the 2001 and 1996 versions. It contains a 25-item [checklist](#) and [flow diagram](#), freely available for viewing and [downloading](#) through this website.



The American Statistician



ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: <http://amstat.tandfonline.com/loi/utas20>

**The ASA's Statement on *p*-Values: Context, Process, and Purpose**

Ronald L. Wasserstein & Nicole A. Lazar



The American Statistician

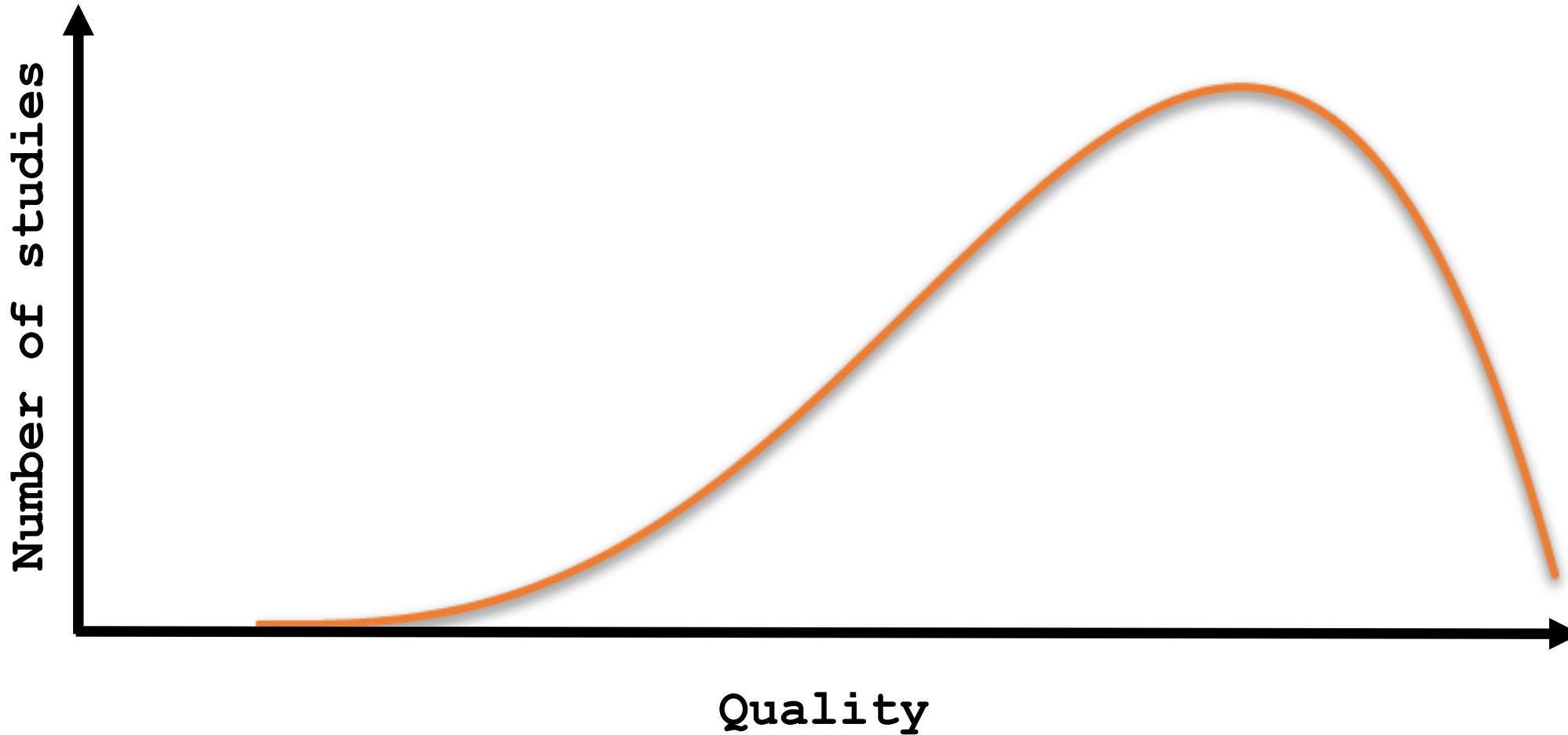


ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: <https://www.tandfonline.com/loi/utas20>

**Moving to a World Beyond "*p* < 0.05"**

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

# Today



# The butterfly has started flapping its wings



Why Plan S [10 Principles](#) Funders & support Implementation About Contact

"After 1 January 2020 scientific publications on the results from research funded by public grants provided by national and European research councils and funding bodies, must be published in compliant Open Access Journals or on compliant Open Access Platforms."



EUROPEAN COMMISSION  
Directorate-General for Research & Innovation

## H2020 Programme

Guidelines on  
FAIR Data Management in Horizon 2020

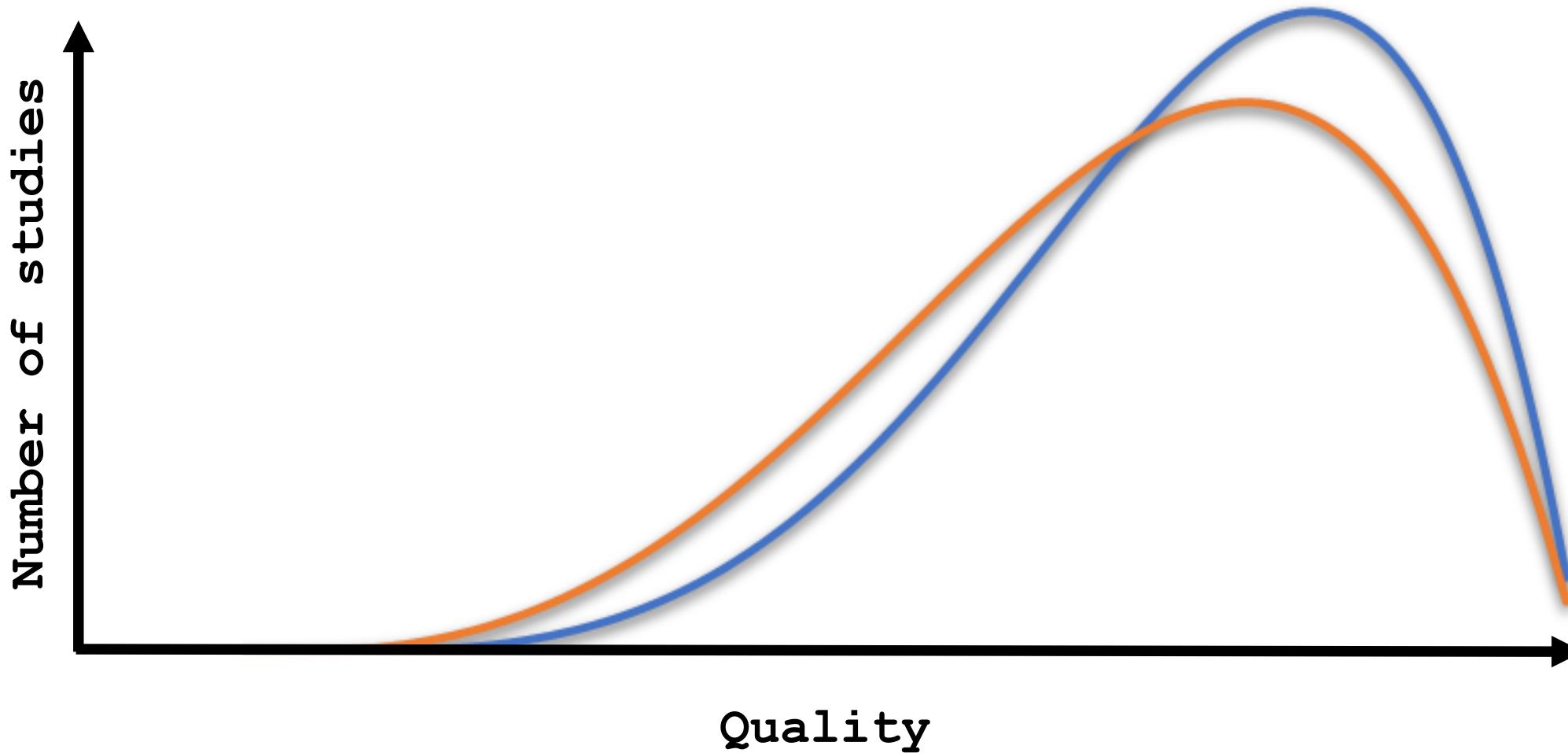


Science Foundation Ireland joins DORA

14th February 2019, Dublin – Science Foundation Ireland has become a signatory to the San Francisco Declaration of Research Assessment (DORA), making a formal commitment to assessing the quality and impact of research through means other than journal impact factors.

The HRB Health Research Board website features a prominent 'Open Research' section. The navigation bar includes links for Funding, Data collections & evidence, Publications, Success stories, News, and About. The 'Open Research' page shows a breadcrumb trail: Home > Funding > Policies and principles > Open Research. A sidebar lists various policy areas like Funding schemes, EU funding support, Manage a grant, etc. The main content area features a circular diagram divided into segments representing different implementation areas: Research data repositories, Open peer review, Academic publishers in research, Line of credit platforms, General and commercial, Journal article open access, Publisher self-archiving, Open access mandates, Researcher self-archiving, and Funder-preprint data sharing. The center of the circle is labeled 'Open Science' with an icon of a person inside a document.

# Tomorrow



# FAIR is a part of your life now!

## Data Guidelines

1. Background
  - 1.1 Open Data Policy
  - 1.2 Fair Data Principles

2. Share Your Data in 3 Steps
  - 2.1 Prepare Your Data for Sharing
  - 2.2 Select a Repository
  - 2.3 Add a Data Availability Statement to Your Manuscript
  - 2.4 Linking your datasets to your article

Some types of data benefit from visualization within the article. Wellcome Open Research welcomes the submission of manuscripts featuring [Plot.ly interactive figures](#) and [Code Ocean compute capsules](#). For further detail, please [contact us](#).



## Research Data Management

Good data governance and stewardship are key components of good research practice. In this regard, Science Foundation Ireland supports that research data should be Findable, Accessible, Interoperable and Reusable (FAIR)\*. Appropriate data management and data sharing are fundamental to all stages of the research process and support high quality, reproducible research. As such, access to research data arising in whole or in part from SFI funding should be as open as possible.



### FAIR Data Management

Describe the approach to data management that will be taken during and after the project, including who will be responsible for data management and data stewardship. The word limit is 500 words.



### Social Research Ethics Committee (SREC) ETHICS APPROVAL FORM

✉ [srec@ucc.ie](mailto:srec@ucc.ie)

<https://www.ucc.ie/en/research/about/ethics/>

<sup>4</sup> Data management should follow the FAIR guiding principles (Findability, Accessibility, Interoperability & Reusability). See, for example, Wilkinson, M. D. et al. (2016) The FAIR Guiding Principles for Scientific Data Management and Stewardship. Full text: <http://www.nature.com/articles/sdata201618>. It is required that all staff and student researchers store those data which are required to replicate research findings, and the information required to enable re-use of data. Details of the UCC policy on research data storage can be found in section 8 of the Code of Research Conduct (2016): <https://www.ucc.ie/en/media/research/researchatucc/documents/UCCCodeofResearchConduct.pdf>. SREC advises against storing research data on non UCC approved cloud-based storage services. Physical data must be stored in a locked cabinet and you must specify who has permission to access this data.



A set of Digital Object Compliance principles that describes the properties of digital objects that enables them to be findable, accessible, interoperable and reproducible (FAIR).

# But what does that mean?



**Naomi Penfold**  
@npscience

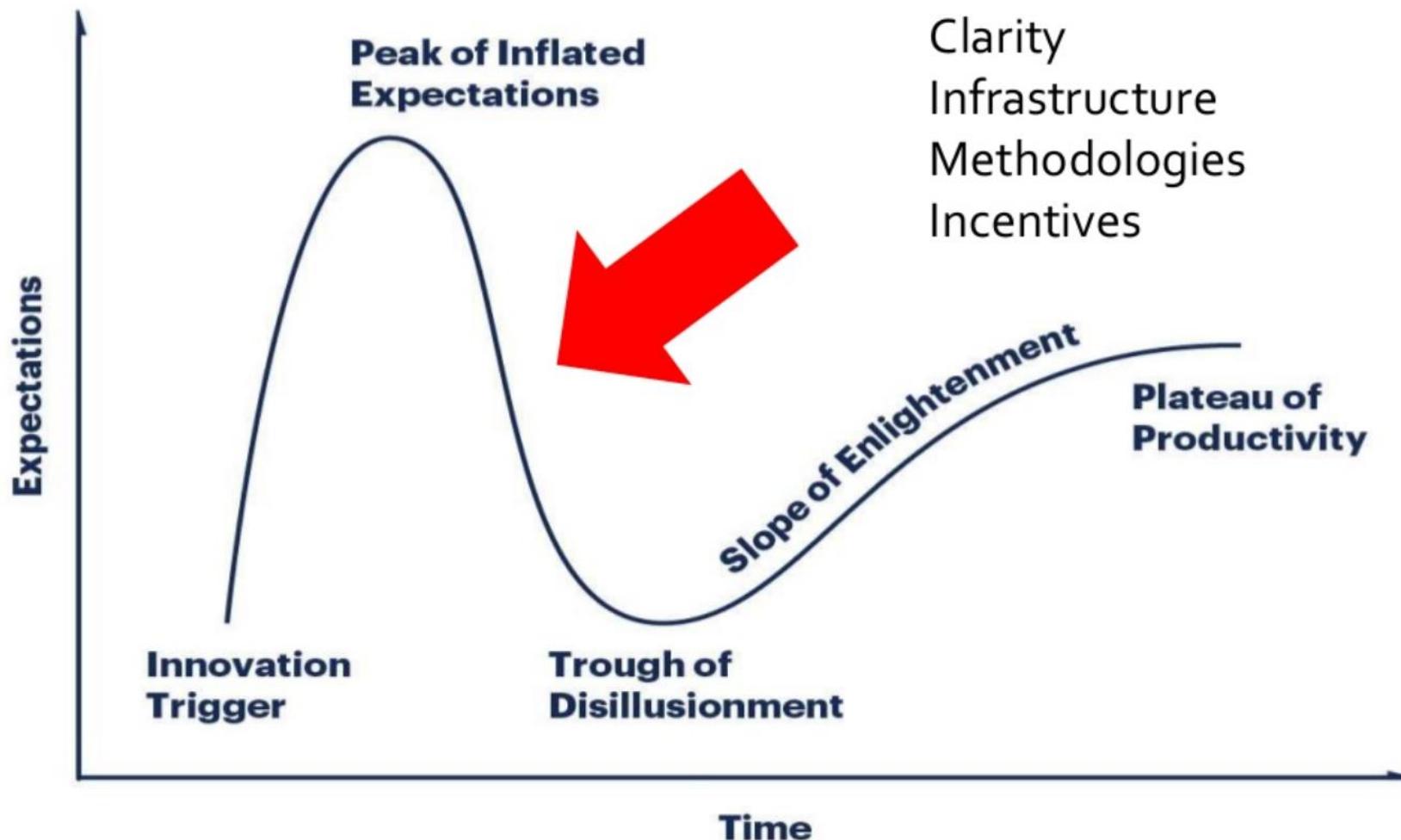


At [#NI2019](#), [@caroleannegoble](#) is getting real about culture change: [#FAIRdata](#) is in the trough of disillusionment and at a point where top-down efforts are considering measuring compliance with a concept that is not (yet, and may never be) concrete or countable

12:57 PM · Sep 1, 2019 · [Twitter Web App](#)

# FAIR Safari – Prof Carole Goble

## The FAIR Hype



# What are the FAIR data principles

The screenshot shows a journal article from the SCIENTIFIC DATA journal. The title of the article is "Comment: The FAIR Guiding Principles for scientific data management and stewardship" by Mark D. Wilkinson et al. The article discusses the need to improve the infrastructure supporting the reuse of scholarly data. It highlights the FAIR Data Principles as a set of guidelines for enhancing data reusability. The article is open access and includes subject categories such as Research data and Publication characteristics. It was received on December 10, 2015, accepted on February 12, 2016, and published on March 15, 2016. The text is presented in a clean, modern layout with a blue header and white background.

www.nature.com/scientificdata

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES

- » Research data
- » Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson et al.\*

Received: 10 December 2015  
Accepted: 12 February 2016  
Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measurable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

**Supporting discovery through good data management**

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of ‘long-term care’ of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes ‘good data management’ is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to ‘data’ in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other’s data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

Correspondence and requests for materials should be addressed to B.M. (email: barend.mons@dtls.nl).  
\*A full list of authors and their affiliations appears at the end of the paper.

- A minimal set of community agreed guiding principles and practices to ensure that research data is:
  - **F**indable
  - **A**ccessible
  - **I**nteroperable
  - **R**eusable
- Initially developed by Dutch Tech Centre for the Life Sciences
- Reviewed and refined through multi-stakeholder practitioner groups, including Force11 and the Research Data Alliance

# What are the FAIR data principles



- F**indable - Assign persistent IDs  
- Machine readable descriptions to support structured searches



- A**ccessible - Retrievable using a standard protocol  
- Metadata available, even if data aren't  
- Authentication and authorization procedure

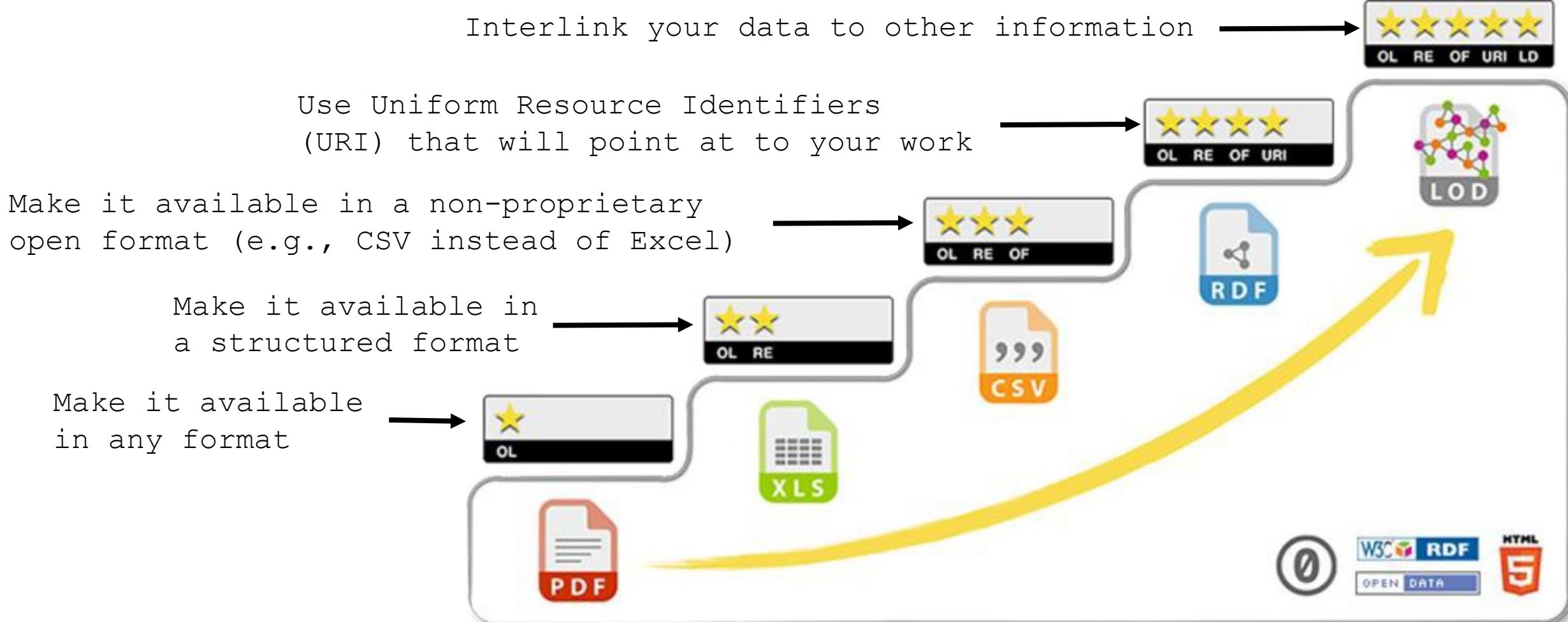


- I**nteroperable - Uses standard (FAIR) vocabularies  
- Linked to other resources



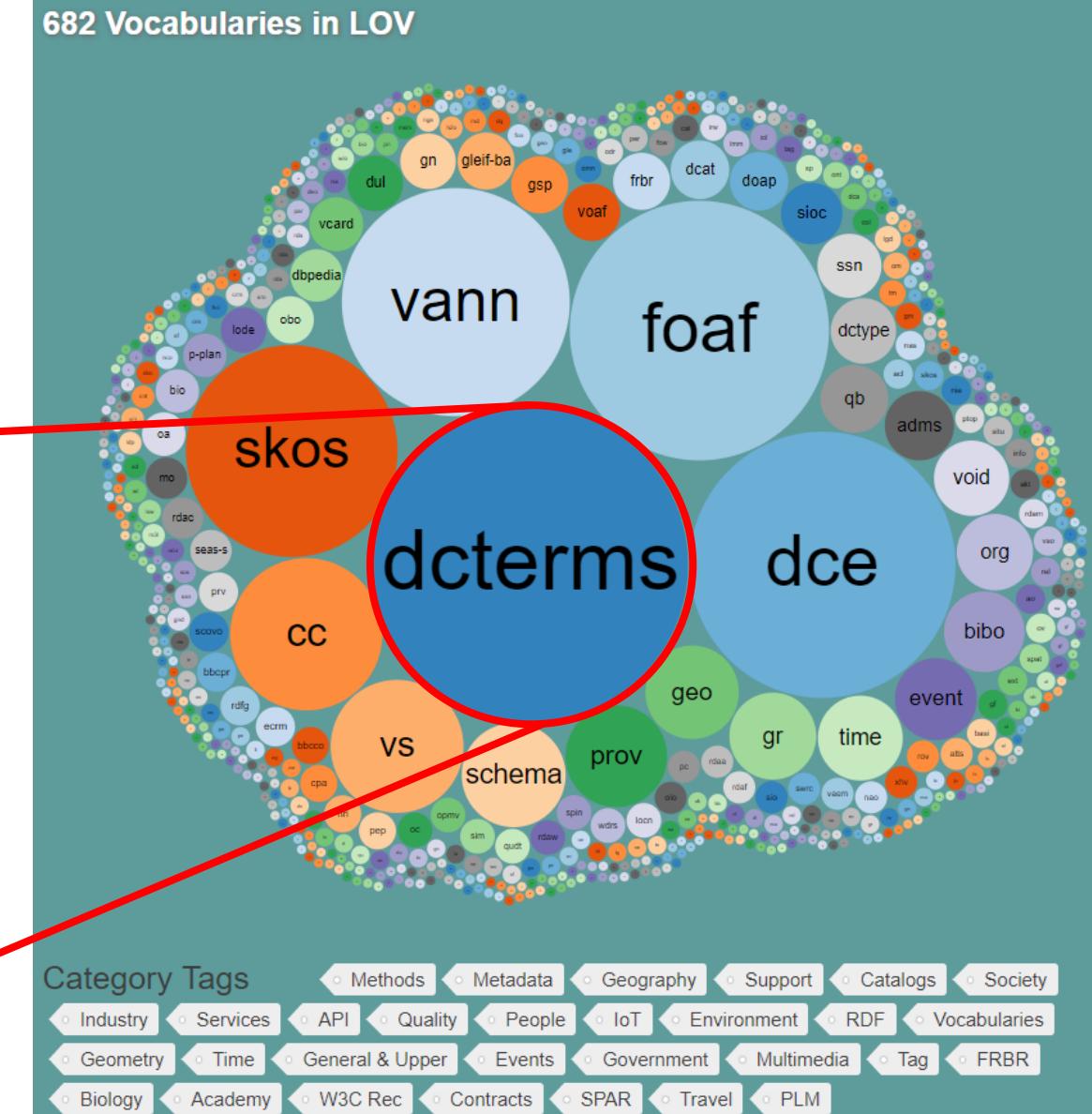
- R**eusable - Clear licenses  
- Provenance  
- Meets domain-relevant community standards

# A path towards FAIR



# Advanced FAIRification – linked vocabularies

The screenshot shows the LOV interface with a navigation bar at the top: VOCABS, TERMS, AGENTS, SPARQL/DUMP. Below is a search bar and a title "Linked Open Vocabularies (LOV)". A red box highlights the "dcterms" vocabulary page. The page includes a "Suggest" button, a "Documentation" link, a "Follow" button, a search bar, and a "Help" icon. The main content area displays the "DCMI Metadata Terms (dcterms)" vocabulary with sections for "Metadata", "Statistics", "Expressivity", "Tags", and "LOD".



# RDF Data

[By input](#)[By URL](#)[By File](#)[By Endpoint](#)

```
1 @prefix schema: <http://schema.org/> .  
2 @prefix : <http://example.org/> .  
3 @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .  
4 @prefix foaf: <http://xmlns.com/foaf/0.1/> .  
5  
6 :alice schema:gender schema:Female ;  
    schema:knows :bob ;  
    schema:name "Alice" .  
9  
10 :bob schema:birthDate "1980-03-10"^^xsd:date ;  
    schema:gender schema:Male ;  
    schema:name "Robert" .  
13  
14 :carol schema:gender schema:Female ;  
    schema:name "Carol" ;  
    foaf:age 23 .
```

# Incorporating FAIR into your routine workflow

F1000



## Your go-to guide to making your data Findable, Accessible, Interoperable, and Reusable (FAIR)

So that you and others can get the most out of your data, it is important that you adhere to the [FAIR principles](#) to ensure your data are **Findable, Accessible, Interoperable, and Reusable** – whilst making your data openly available where it is safe to do so. This is no small task, so here are some ideas to help you get started:

1

### Start with a management plan

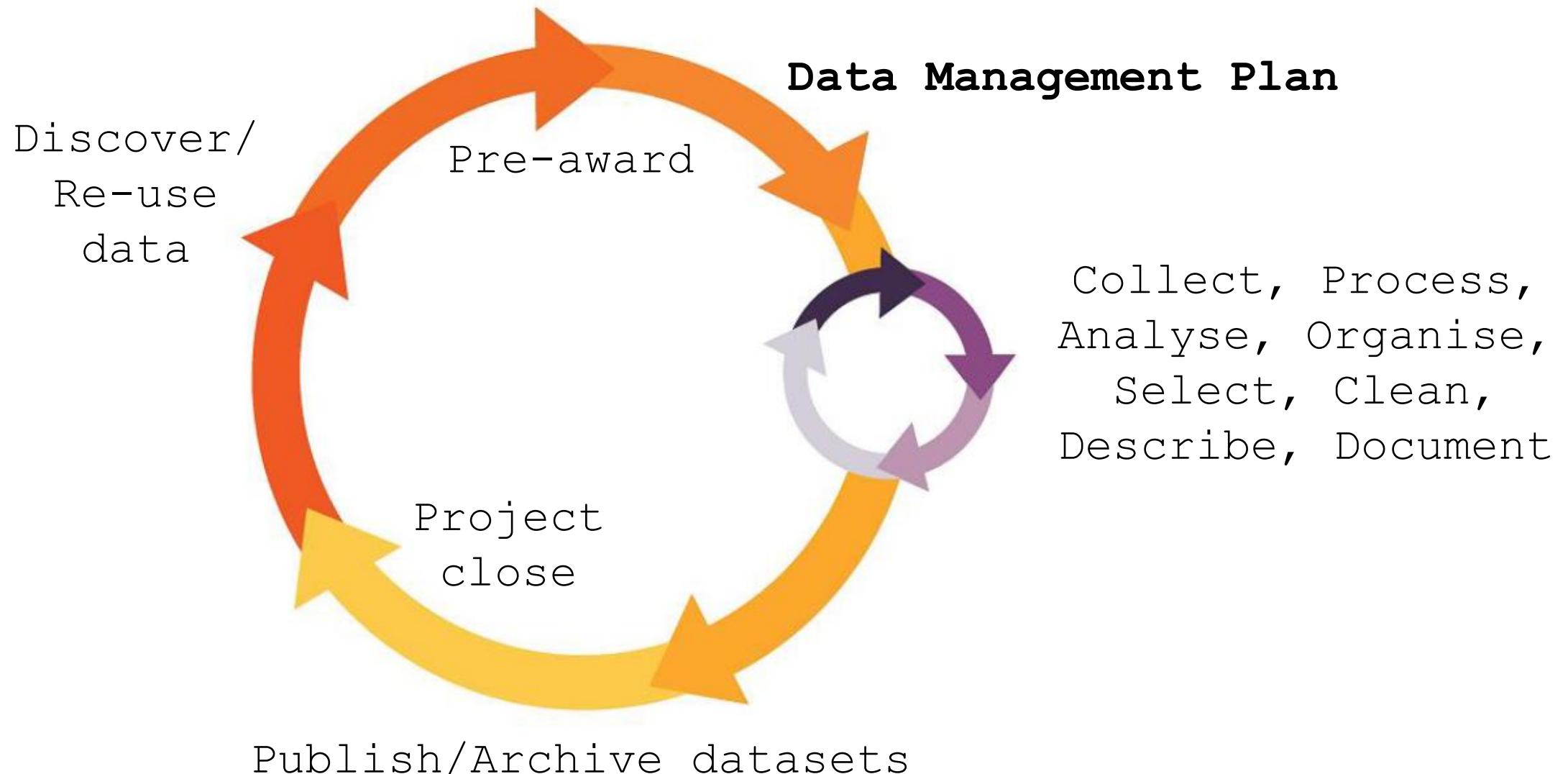
An output management plan (OMP) is a useful starting point for collecting or creating data, software, research materials, and intellectual property. Creating an OMP before you begin your research, and updating it throughout the research cycle, will help ensure that your outputs are as open and **FAIR** as possible when your project is complete.

Some funders require grant-holders to produce a plan as part of their application for funding, and/or after funding has been secured.

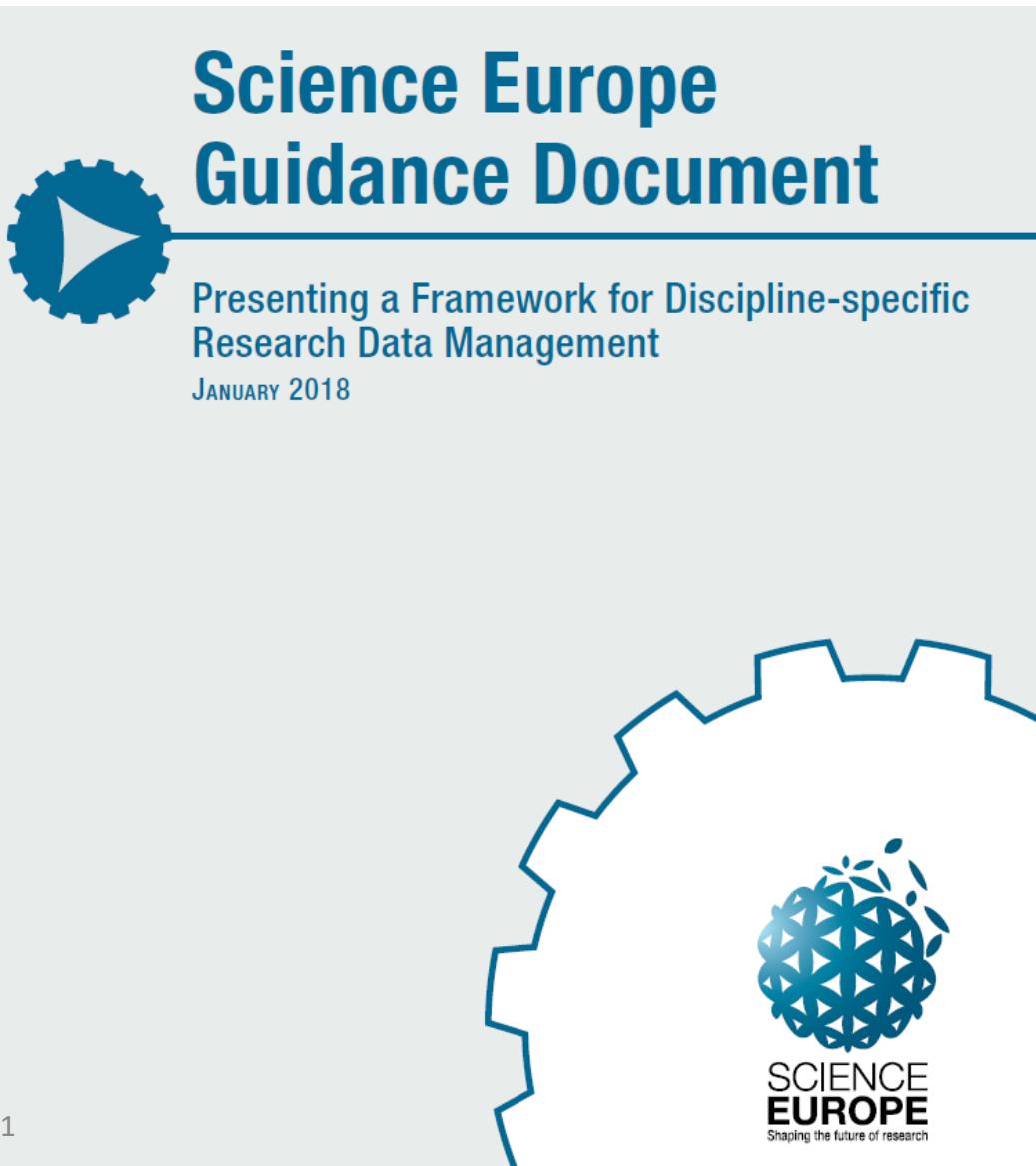
You should consider:

- What outputs you will be creating or collecting, and how these will be documented
- What ethical or legal requirements, if any, apply to the outputs
- How you will organise, store, secure, and share the outputs
- What resources are required and who is responsible

# Incorporating FAIR into your routine workflow



# It all starts with a Data Management Plan



**Seven main headings:**

1. Data collection
2. Documentation and Meta-data
3. Ethics and Legal Compliance
4. Storage and Backup
5. Selection and Preservation
6. Data Sharing
7. Responsibilities and Resources

# Taking those first steps

**DMPONLINE**

Home    Public DMPs    Funder requirements    Help

Language ▾

## Welcome

DMPonline helps you to create, review, and share data management plans that meet institutional and funder requirements. It is provided by the Digital Curation Centre (DCC).

Join the growing international community that have adopted DMPonline:



17,622 Users



203 Organisations



23,083 Plans



89 Countries

Some funders mandate the use of DMPonline, while others point to it as a useful option. You can [download funder templates](#) without logging in, but the tool provides tailored guidance and example answers from the DCC and many research organisations. Why not sign up for an account and try it out?

Sign in    Create account

\* Email

\* Password

Forgot password?

Remember email

Sign in

- OR -

Sign in with your institutional credentials

University College Cork

# An amazing new treatment that will cure all that ails us

Project Details

Plan overview

Write Plan

Share

Download

[expand all](#) | [collapse all](#)

0/11 answered

## Data and software outputs (0 / 6)

The data and software outputs your research will generate



Guidance

Comments

Wellcome Trust

DCC

Consider and briefly describe:

- the types of data and software the proposed research will generate
- which data and software will have value to other research users and could be shared
- the formats and quality standards that will

# Taking those first steps



Smart Data Management Plans for FAIR Open Science  
For Serious Researchers and Data Stewards

## How to use the Data Stewardship Wizard

We offer several options how to use the Data Stewardship Wizard, each suited for different use case.

Demo Instance	Researchers Instance	Self-hosted instance	Instance hosted by us
<p><i>For exploring the DSW features</i></p> <ul style="list-style-type: none"><li>• Easy to sign up and use</li><li>• A shared instance with other users</li><li>• <b>Not for serious usage</b></li></ul>	<p><i>For individual researchers</i></p> <ul style="list-style-type: none"><li>• Easy to sign up and use</li><li>• Ready to use Knowledge Models</li><li>• Privacy and stability</li></ul>	<p><i>For organizations</i></p> <ul style="list-style-type: none"><li>• All the DSW features available</li><li>• Your own instance</li><li>• You need to host and run the instance by yourself</li></ul>	<p><i>For organizations</i></p> <p>We offer managing the DS Wizard instance for interesting projects that want to use it seriously but don't want to run it by themselves.</p>

# Exploring standards in your field



A curated, informative and educational resource on data and metadata *standards*, inter-related to *databases* and *data policies*.

## HOW CAN WE HELP?

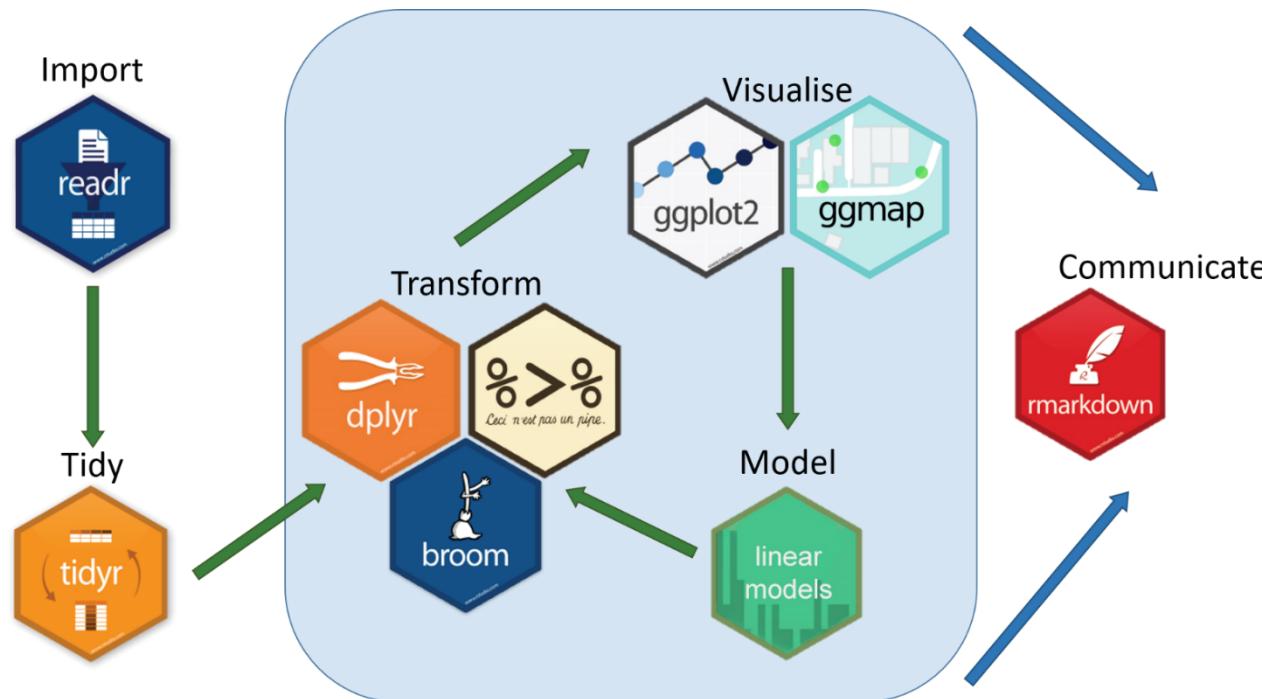
We guide consumers to discover, select and use these resources with confidence, and producers to make their resource more discoverable, more widely adopted and cited.



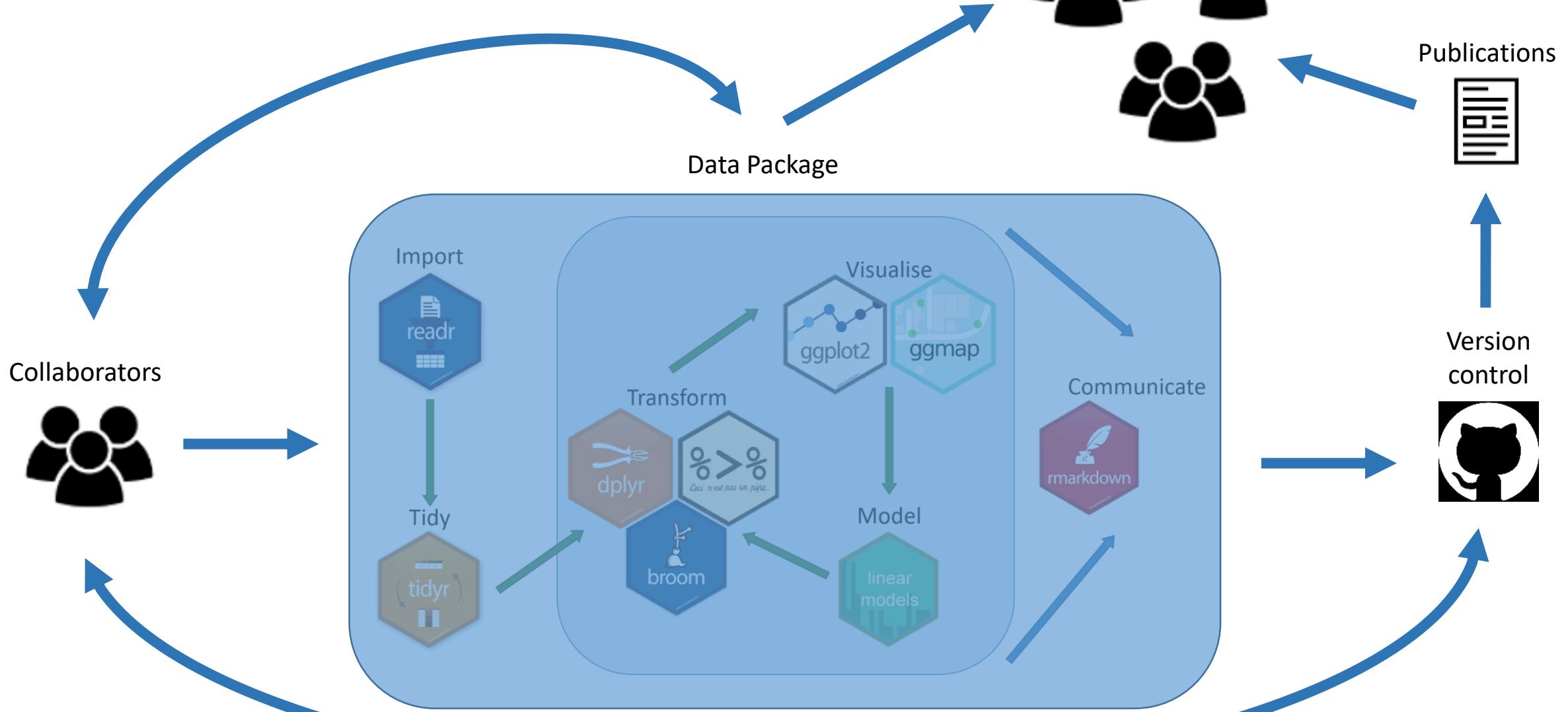
### Research data facilitators, librarians, trainers

Use FAIRsharing to provide a foundation on which to create or enrich educational lectures, training and teaching material, and to plug into data management planning tools...  
[\[read more\]](#)

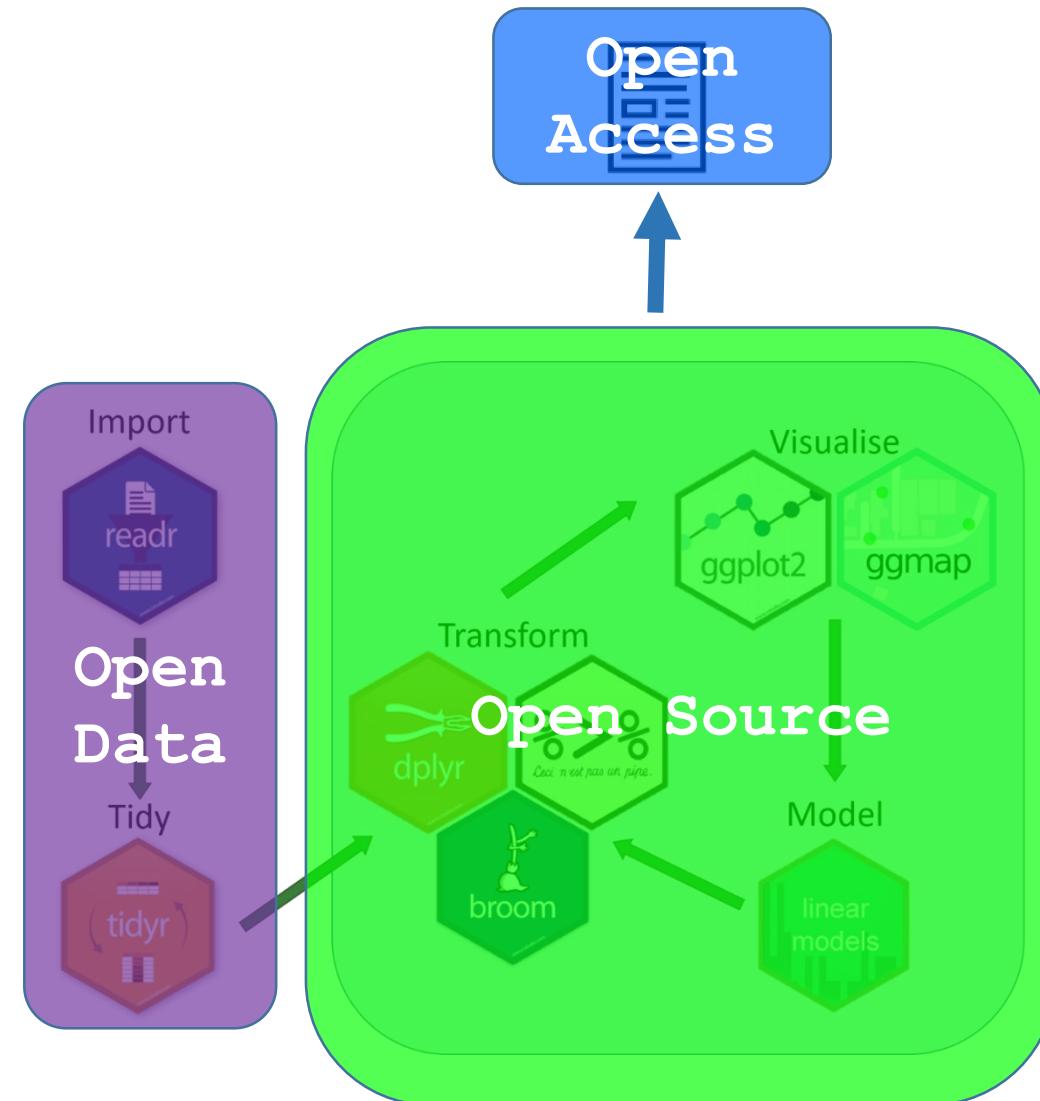
# Putting the pieces together using R



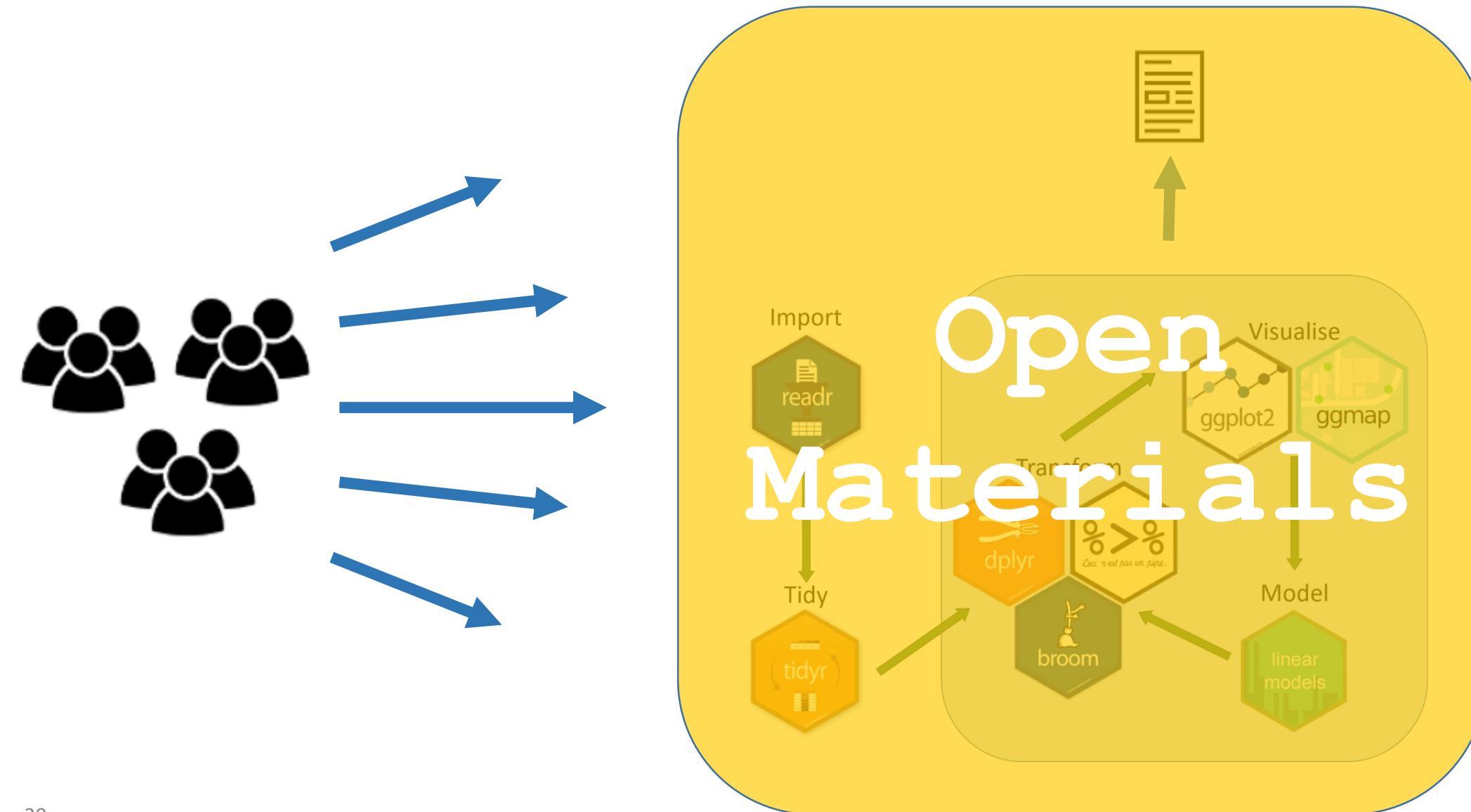
# The bigger picture



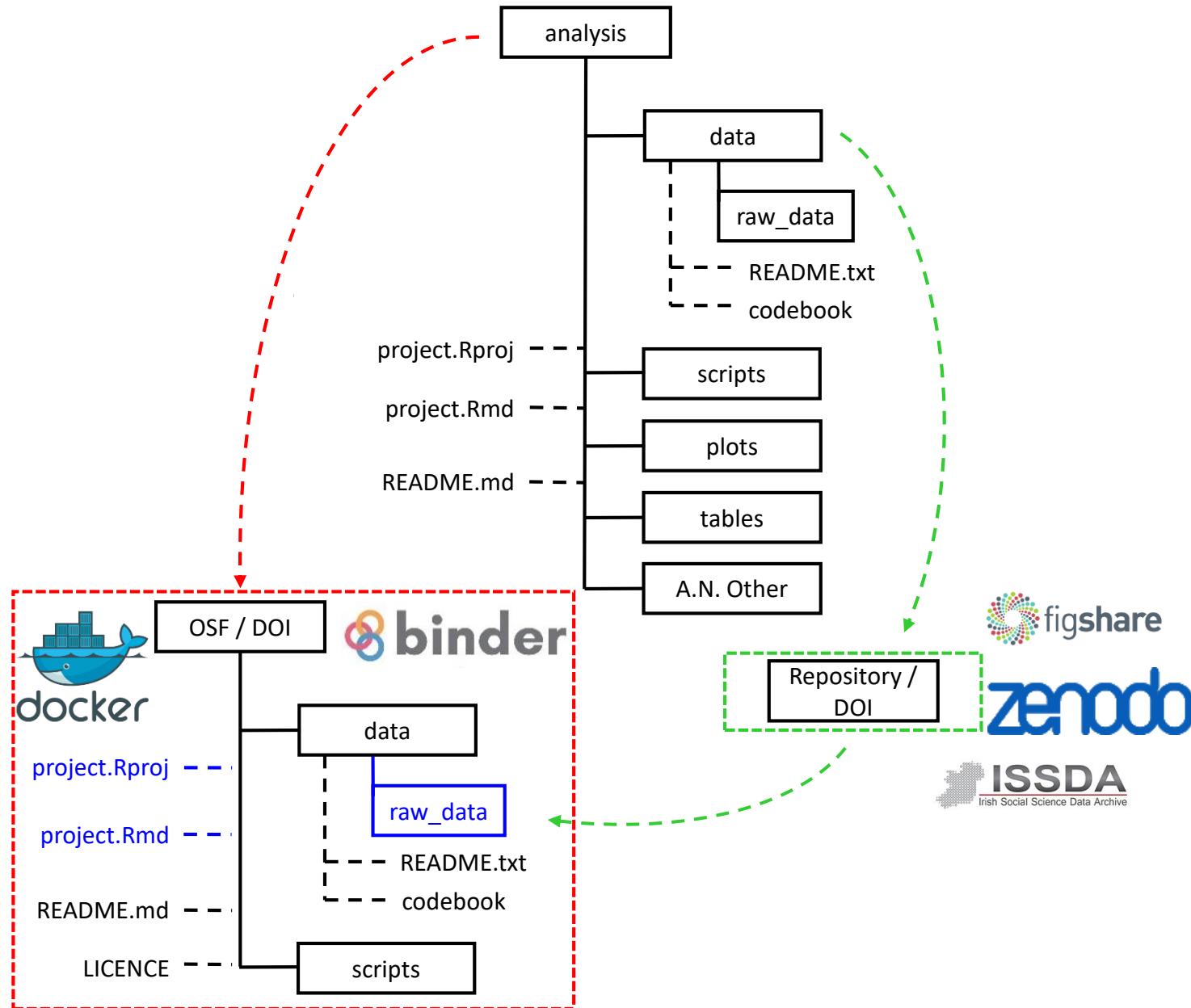
# The ‘Open Science’ picture



# The ‘Open Science’ picture



# What does this allow us to do?



# Project Packaging



Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

Build and launch a repository

GitHub repository name or URL

GitHub ▾

Git branch, tag, or commit

Path to a notebook file (optional)

File ▾

launch

# Our real life experiment



- UV light has potential to change the secondary metabolite composition (colour) of bronze/red lettuce
- Experimental setup:
  - 3 lettuce varieties
  - 3 UV filter conditions
  - 3 weeks duration

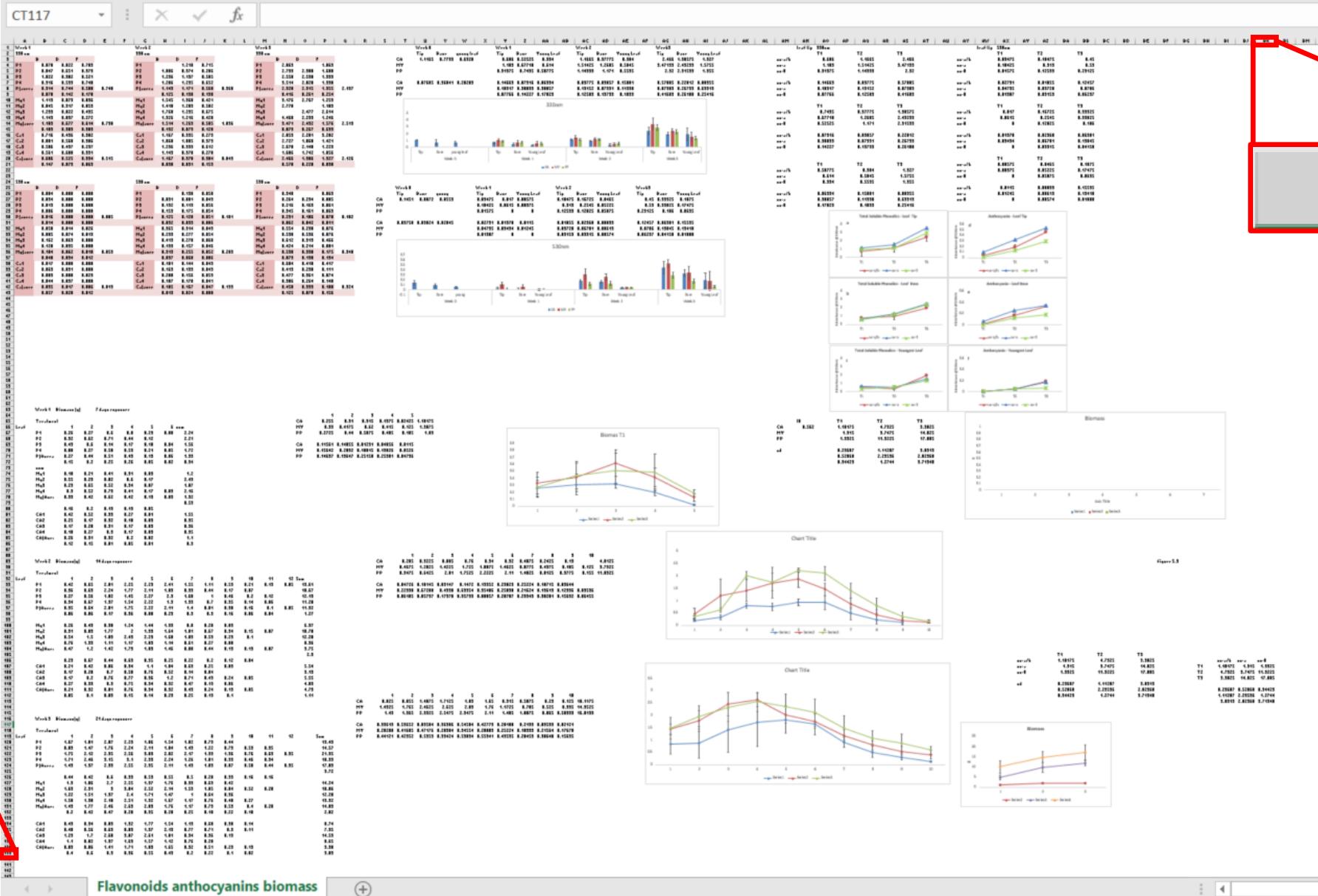
# Real data comes with real problems

Raw Data wk 1-3 Lettuce Exp 1 - Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Week 1						Week 2						Week 3					
2	330 nm						330 nm						330 nm					
3		B	D	F				B	D	F				B	D	F		
4	P1	0.870	0.822	0.703			P1						1	2.869		1.069		
5	P2	0.847	0.651	0.379			P2						2	2.739	2.380	1.688		
6	P3	1.022	0.902	0.521			P3	1.236	1.197	0.585			P3	2.558	2.538	1.333		
7	P4	0.916	0.599	0.748			P4	1.206	1.295	0.652			P4	3.514	2.028	1.330		
8	P(average)	0.914	0.744	0.588	0.748		P(average)	1.149	1.171	0.560	0.960		P(average)	2.920	2.315	1.355	2.197	
9		0.078	0.142	0.170				0.125	0.138	0.190				0.416	0.261	0.254		
10	My1	1.119	0.873	0.896			My1	1.545	1.360	0.421			My1	3.176	2.767	1.259		
11	My2	0.845	0.917	0.853			My2	1.418	1.203	0.502			My2	2.778		1.183		
12	My3	1.299	0.822	0.435			My3	1.768	1.295	0.675			My3		2.477	2.614		
13	My4	1.149	0.097	0.272			My4	1.326	1.216	0.420			My4	4.460	2.233	1.246		
14	My(average)	1.103	0.677	0.614	0.798		My(average)	1.514	1.269	0.505	1.096		My(average)	3.471	2.492	1.576	2.513	
15		0.189	0.389	0.309				0.192	0.073	0.120				0.879	0.267	0.693		
16	Ca1	0.716	0.496	0.382			Ca1	1.167	0.935	0.273			Ca1	2.853	2.201	3.202		
17	Ca2	0.881	0.568	0.386			Ca2	1.060	1.005	0.373			Ca2	2.727	1.860	1.421		
18	Ca3	0.586	0.437	0.237			Ca3	1.296	0.993	0.612			Ca3	2.678	2.140	1.229		
19	Ca4	0.561	0.600	0.331			Ca4	1.143	0.978	0.278			Ca4	1.606	1.742	1.856		
20	Ca(average)	0.686	0.525	0.334	0.515		Ca(average)	1.167	0.978	0.384	0.843		Ca(average)	2.466	1.986	1.927	2.126	
21		0.147	0.073	0.069				0.098	0.031	0.159				0.578	0.220	0.890		
22																		
23																		
24	530 nm						530 nm						530 nm					
25		B	D	F				B	D	F				B	D	F		
26	P1	0.004	0.000	0.000			P1		0.138	0.050				P1	0.340		0.069	
27	P2	0.034	0.000	0.000			P2		0.091	0.081	0.043			P2	0.264	0.234	0.085	CA
28	P3	0.019	0.000	0.000			P3		0.132	0.119	0.056			P3	0.216	0.163	0.061	MY

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do...

Normal Page Break Preview Custom Layout Views Workbook Views Show Zoom 100% Zoom to Selection Window New Arrange Freeze All Panes Hide Synchronous Scrolling Reset Window Position Window Switch Windows Macros Macros



# Take small steps to enact big changes

THE AMERICAN STATISTICIAN  
2018, VOL. 72, NO. 1, 2–10  
<https://doi.org/10.1080/00031305.2017.1375989>



OPEN ACCESS



## Data Organization in Spreadsheets

Karl W. Broman<sup>a</sup> and Kara H. Woo<sup>b</sup>

<sup>a</sup>Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; <sup>b</sup>Information School, University of Washington, Seattle, WA

### ABSTRACT

Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this article offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, do not include calculations in the raw data files, do not use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.

### ARTICLE HISTORY

Received June 2017  
Revised August 2017

### KEYWORDS

Data management; Data organization; Microsoft Excel; Spreadsheets

# Less stress, more success

1	A	B	C	D	E	F	G	H	I	J	K	L
1	id	week_no	filter_nam	treatment	replicate_no	flavonoids	biomass	variety	date	investigator		
2	1	0	ptp	nofilter	1	1.061	0.39	cos	2019/04/01	Darren Dahly		
3	2	0	ptp	nofilter	2	1.1805	0.42	cos	2019/04/01	Darren Dahly		
4	3	0	ptp	nofilter	3	1.0345	0.62	cos	2019/04/01	Darren Dahly		
5	4	0	ptp	nofilter	4	1.094	0.63	cos	2019/04/01	Brendan Palmer		
6	1	0	my	nofilter	1	1.061	0.39	cos	2019/04/01	Brendan Palmer		
7	2	0	my	nofilter	2	1.1805	0.42	cos	2019/04/01	Brendan Palmer		
8	3	0	my	nofilter	3	1.0345	0.62	cos	2019/04/01	Brendan Palmer		
9	4	0	my	nofilter	4	1.094	0.63	cos	2019/04/01	Brendan Palmer		
10	1	0	ca	nofilter	1	1.061	0.39	cos	2019/04/01	Brendan Palmer		
11	2	0	ca	nofilter	2	1.1805	0.42	cos	2019/04/01	Brendan Palmer		
12	3	0	ca	nofilter	3	1.0345	0.62	cos	2019/04/01	Brendan Palmer		
13	4	0	ca	nofilter	4	1.094	0.63	cos	2019/04/01	Darren Dahly		
14	5	1	ptp	filter	1	0.87	0.76	cos	2019/04/08	Darren Dahly		
15	6	1	ptp	filter	2	0.847	0.95	cos	2019/04/08	Darren Dahly		
16	7	1	ptp	filter	3	1.022	0.95	cos	2019/04/08	Darren Dahly		
17	8	1	ptp	filter	4	0.916	0.95	cos	2019/04/08	Darren Dahly		
18	9	1	my	filter	1	1.119	1.55	cos	2019/04/08	Darren Dahly		
19	10	1	my	filter	2	0.845	3.16	cos	2019/04/08	Darren Dahly		
20	11	1	my	filter	3	1.299	4.9	cos	2019/04/08	Brendan Palmer		
21	12	1	my	filter	4	1.149	5.5	cos	2019/04/08	Brendan Palmer		
22	13	1	ca	filter	1	0.716	5.5	cos	2019/04/08	Brendan Palmer		
23	14	1	ca	filter	2	0.881	7.94	cos	2019/04/08	Brendan Palmer		
24	15	1	ca	filter	3	0.586	8.71	cos	2019/04/08	Brendan Palmer		
25	16	1	ca	filter	4	0.561	8.71	cos	2019/04/08	Brendan Palmer		
26	17	2	ptp	filter	1	0	14.45	cos	2019/04/15	Brendan Palmer		
27	18	2	ptp	filter	2	1.006	2.14	cos	2019/04/15	Brendan Palmer		
28	19	2	ptp	filter	3	1.236	1.86	cos	2019/04/15	Brendan Palmer		
29	20	2	ptp	filter	4	1.206	1.2	cos	2019/04/15	Brendan Palmer		
30	21	2	mv	filter	1	1.545	2.45	cos	2019/04/15	Brendan Palmer		

data

dictionary

values



# Less stress, more success

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	week_no	filter_name	treatment	replicate_no	flavonoids	biomass	variety	date	investigator		
2	1	0	ptp	nofilter	1	1.061	0.39	cos	2019/04/01	Darren Dahly		
3	2	0	ptp		A	B	C	D	E			
4	3	0	ptp		1	field_name	data_type	data_format	example	standard_units	description	
5	4	0	ptp		2	id	numeric	integer		23	NA	Unique identifier applied to each observation
6	1	0	my		3	week_no	numeric	integer		1	NA	Week number, 1 = 7 days exposure, 2 = 14 days exposure
7	2	0	my		4	filter_name	character	NA	my	NA		3 filter types; 'ptp' = polytunnel plastic blocks all UV light
8	3	0	my		5	treatment	character	NA	filter	NA		Presence or absence of a filter at the time of sampling
9	4	0	my		6	replicate_no	numeric	integer		1	NA	The number of replicates in each treatment
10	1	0	ca		7	flavonoids	numeric	double	0.3421	parts per million (ppm)		Leaf disc taken from the tip of the most mature leaf at th
11	2	0	ca		8	biomass	numeric	double		gram (g)		Above ground biomass on the day of harvest
12	3	0	ca		9	variety	character	NA	cos	NA		3 commerical varieties of red lettuce used; 'cos' = Cos Di
13	4	0	ca		10	date	date	YYYY/MM/DD	2019/06/28	ISO 8601		Experiment date
14	5	1	ptp		11	investigator	character	Firstname Lastname	Aoife Coffey	NA		Primary researcher who performed the experiment
15	6	1	ptp		12							
16	7	1	ptp		13							
17	8	1	ptp		14							
18	9	1	my		15							
19	10	1	my		16							
20	11	1	my		17							
21	12	1	my		18							
22	13	1	ca		19							
23	14	1	ca		20							
24	15	1	ca		21							
25	16	1	ca		22							
26	17	2	ptp		23							
27	18	2	ptp		24							
28	19	2	ptp		25							
29	20	2	ptp		26							
30	21	2	mv		27							
			dictionary		28							
					29							
					30							

# Less stress, more success

The screenshot illustrates a data entry interface with two main tabs: "data" and "dictionary".

**Data Tab:**

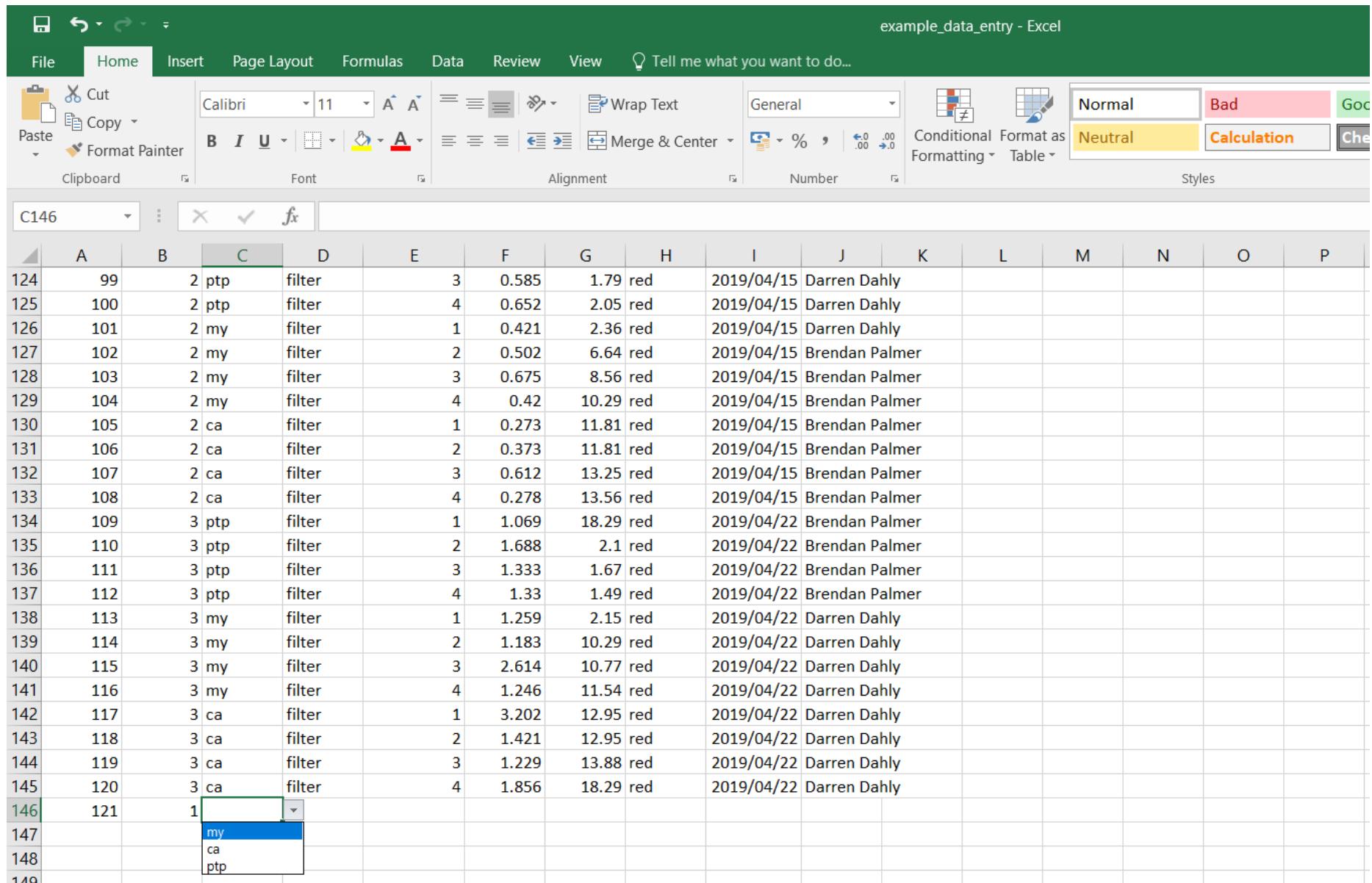
	A	B	C	D	E	F	G	H	I	J	K	L
1	id	week_no	filter_name	treatment	replicate_no	flavonoids	biomass	variety	date	investigator		
2	1	0	ptp	nofilter	1	1.061	0.39	cos	2019/04/01	Darren Dahly		
3	2	0	ptp									
4	3	0	ptp									
5	4	0	ptp									
6	1	0	my									
7	2	0	my									
8	3	0	my									
9	4	0	my									
10	1	0	ca									
11	2	0	ca									
12	3	0	ca									
13	4	0	ca									
14	5	1	ptp									
15	6	1	ptp									
16	7	1	ptp									
17	8	1	ptp									
18	9	1	my									
19	10	1	my									
20	11	1	my									
21	12	1	my									
22	13	1	ca									
23	14	1	ca									
24	15	1	ca									
25	16	1	ca									
26	17	2	ptp									
27	18	2	ptp									
28	19	2	ptp									
29	20	2	ptp									
30	21	2	mv									

**Dictionary Tab:**

	A	B	C	D	E	F	G	H	I	J	K
1	field_name	data_type	data_format	example	standard_units	description					
2	id	numeric	integer			Unique identifier applied to each observation.					
3	week_no	numeric	integer								
4	filter_name	character	NA								
5	treatment	character	NA								
6	replicate_no	numeric	integer								
7	flavonoids	numeric	double								
8	biomass	numeric	double								
9	variety	character	NA								
10	date	date	YYYY/MM/DD								
11	investigator	character	Firstname Lastname								

The "data" tab displays the experimental data, while the "dictionary" tab provides the schema definition for the fields.

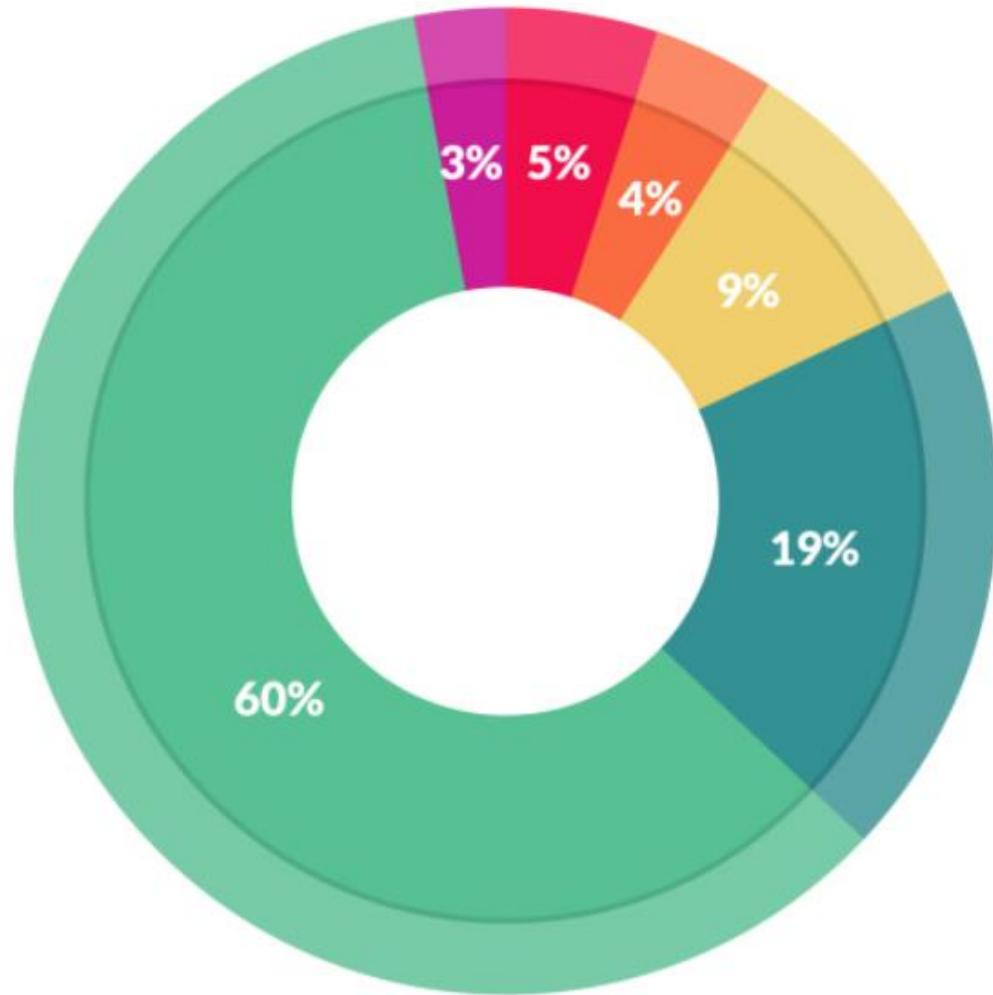
# Less stress, more success



The screenshot shows a Microsoft Excel spreadsheet titled "example\_data\_entry - Excel". The ribbon is visible at the top, with the "Home" tab selected. The main area displays a table of data with columns A through P. Row 146 is currently selected, and a dropdown menu is open over cell C146, showing options: "my", "ca", and "ptp". The data in the table includes numerical values, text entries like "filter", and dates/times.

C146	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
124	99	2	ptp	filter		3	0.585	1.79	red	2019/04/15	Darren Dahly					
125	100	2	ptp	filter		4	0.652	2.05	red	2019/04/15	Darren Dahly					
126	101	2	my	filter		1	0.421	2.36	red	2019/04/15	Darren Dahly					
127	102	2	my	filter		2	0.502	6.64	red	2019/04/15	Brendan Palmer					
128	103	2	my	filter		3	0.675	8.56	red	2019/04/15	Brendan Palmer					
129	104	2	my	filter		4	0.42	10.29	red	2019/04/15	Brendan Palmer					
130	105	2	ca	filter		1	0.273	11.81	red	2019/04/15	Brendan Palmer					
131	106	2	ca	filter		2	0.373	11.81	red	2019/04/15	Brendan Palmer					
132	107	2	ca	filter		3	0.612	13.25	red	2019/04/15	Brendan Palmer					
133	108	2	ca	filter		4	0.278	13.56	red	2019/04/15	Brendan Palmer					
134	109	3	ptp	filter		1	1.069	18.29	red	2019/04/22	Brendan Palmer					
135	110	3	ptp	filter		2	1.688	2.1	red	2019/04/22	Brendan Palmer					
136	111	3	ptp	filter		3	1.333	1.67	red	2019/04/22	Brendan Palmer					
137	112	3	ptp	filter		4	1.33	1.49	red	2019/04/22	Brendan Palmer					
138	113	3	my	filter		1	1.259	2.15	red	2019/04/22	Darren Dahly					
139	114	3	my	filter		2	1.183	10.29	red	2019/04/22	Darren Dahly					
140	115	3	my	filter		3	2.614	10.77	red	2019/04/22	Darren Dahly					
141	116	3	my	filter		4	1.246	11.54	red	2019/04/22	Darren Dahly					
142	117	3	ca	filter		1	3.202	12.95	red	2019/04/22	Darren Dahly					
143	118	3	ca	filter		2	1.421	12.95	red	2019/04/22	Darren Dahly					
144	119	3	ca	filter		3	1.229	13.88	red	2019/04/22	Darren Dahly					
145	120	3	ca	filter		4	1.856	18.29	red	2019/04/22	Darren Dahly					
146	121	1	my ca ptp													
147																
148																
149																

# Resources are being wasted by not doing this



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

# Try it out yourself

Search or jump to... / Pull requests Issues Marketplace Explore



Overview Repositories 14 Projects 0 Stars 1 Followers 12 Following 10

Pinned Order updated. Customize your pins

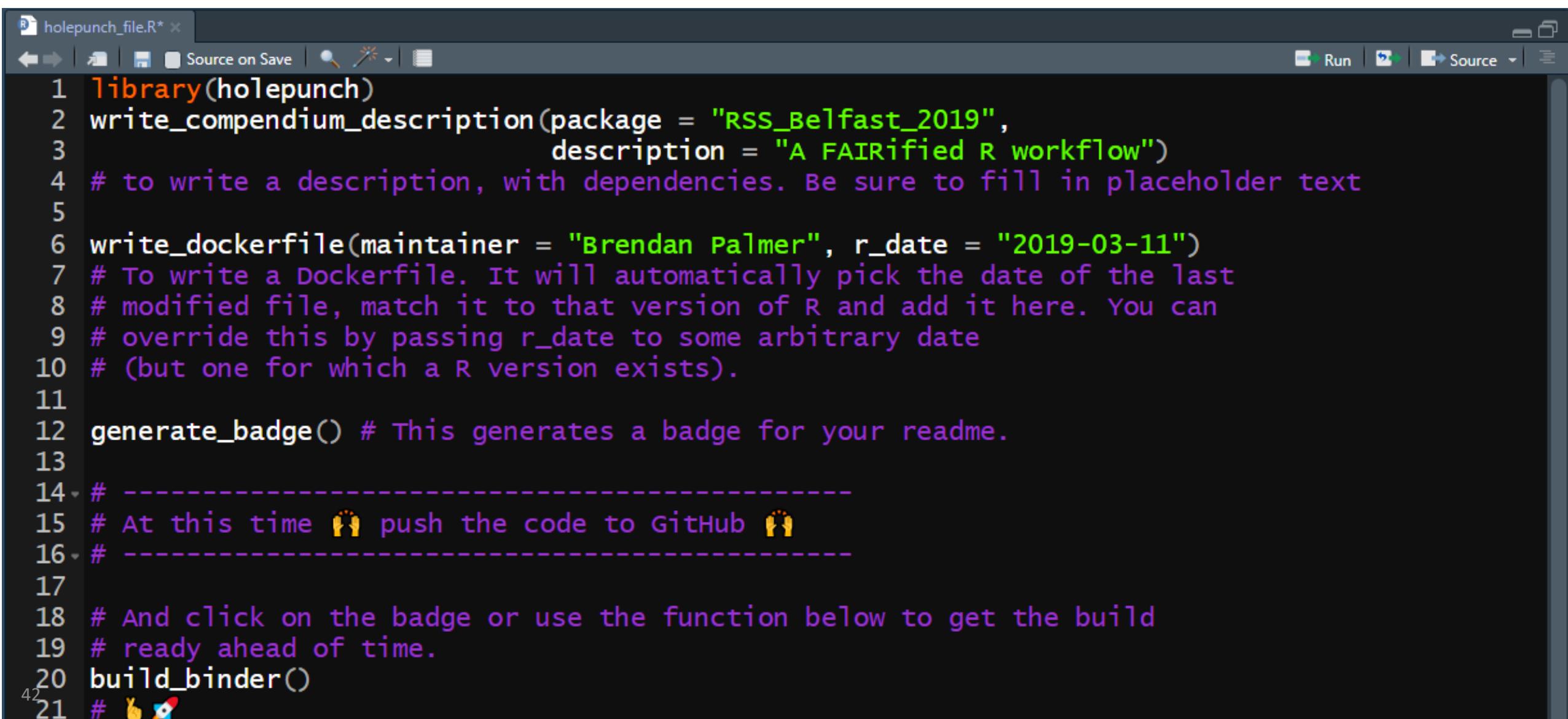
 <a href="#">RSS_Belfast_2019</a> Data FAIRification using R/RStudio workflows  ● R	 <a href="#">R-A_Hitchhikers_Guide_to_Reproducible_Research</a> A 3-day R course given in University College Cork that encompasses various elements off reproducible research facilitated through RStudio projects, the R tidyverse language and reporting using R Ma...  ● HTML ★ 2
 <a href="#">RCR</a> Section of the UCC Reproducible Conduct of Research digital badge dedicated to exposing researchers to reproducible research practices.  ● HTML	 <a href="#">lunchtime_sessions</a> Short 1 hour introductions to R-related topics such as creating R projects, using GitHub through RStudio and more  ● HTML ★ 1

Set status

**Brendan Palmer**  
bapalmer

Edit profile

# It doesn't get much easier than this!



A screenshot of an RStudio interface showing an R script named 'holepunch\_file.R'. The code uses the 'holepunch' package to generate a Dockerfile, a badge, and a Compendium description. It includes comments explaining the purpose of each step and how to customize it.

```
holepunch_file.R* | Run | Source | Source |
```

```
1 library(holepunch)
2 write_compendium_description(package = "RSS_Belfast_2019",
3                               description = "A FAIRified R workflow")
4 # to write a description, with dependencies. Be sure to fill in placeholder text
5
6 write_dockerfile(maintainer = "Brendan Palmer", r_date = "2019-03-11")
7 # To write a Dockerfile. It will automatically pick the date of the last
8 # modified file, match it to that version of R and add it here. You can
9 # override this by passing r_date to some arbitrary date
10 # (but one for which a R version exists).
11
12 generate_badge() # This generates a badge for your readme.
13
14 # -----
15 # At this time 🚀 push the code to GitHub 🚀
16 # -----
17
18 # And click on the badge or use the function below to get the build
19 # ready ahead of time.
20 build_binder()
21 # 🚀
```

# Hello... Is anyone there?...



**bioRxiv**  
THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | ALERTS / RSS | CHANNELS

Search



Advanced Search

Contradictory Results

2 comments

Previous

Next

## Revisiting the decay of scientific email addresses

Posted May 12, 2019.

Raul Rodriguez-Esteban, Dina Vishnyakova, Fabio Rinaldi

doi: <https://doi.org/10.1101/633255>

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract

Full Text

Info/History

Metrics

Preview PDF

Tweet

Like 7

Download PDF

Data/Code

Email

Share

Citation Tools

### Abstract

Email is the primary means of communication for scientists. However, scientific authors change email address over time. Using a new method, we have calculated that approximately 18% of all authors' contact email addresses in MEDLINE are invalid. While an unfortunate number, it is, however, lower than previously estimated. To mitigate this problem, institutions should provide email forwarding and scientific authors should use more stable email addresses. In fact, a steadily growing share already use free private email addresses: 32% of all new addresses in MEDLINE in 2018 were of this kind.

### Subject Area

Scientific Communication and Education

### Subject Areas

#### All Articles

Animal Behavior and Cognition

Biochemistry

Bioengineering

Bioinformatics

# Can I see your data an code?

1989

\* Corresponding author.

1999

\* Corresponding author. Mailing address: Institute of Human Virology, 725 West Lombard St., Rm. N649, University of Maryland, Baltimore, MD 21201. Phone: (410) 706-4680. Fax: (410) 706-4694. E-mail: devico@umbi.umd.edu.

2009

\* Corresponding author. Mailing address: Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 201 Althouse Laboratory, University Park, PA 16802. Phone: (814) 863-8705. Fax: (814) 865-7927. E-mail: cec9@psu.edu.

Also 2019



The screenshot shows the homepage of the Journal of Virology. It features the journal's logo, "AMERICAN SOCIETY FOR MICROBIOLOGY", and the title "Journal of Virology". There is a search bar and an "Advanced Search" link. A horizontal menu includes "Home", "Articles", "For Authors", "About the Journal", and "Subscribe". Below the menu, a link to "Genetic Diversity and Evolution | Spotlight" is visible.

## Single-Cell Virus Sequencing of Influenza Infections That Trigger Innate Immunity

Finally, we process the annotated cell-gene matrix in R to generate the plots shown in this paper. This analysis utilized a variety of R and Bioconductor ([90](#)) packages, including Monocle ([91](#), [92](#)) and ggplot2. A Jupyter notebook that performs these analyses is at

[https://github.com/jbloomlab/IFNsorted\\_flu\\_single\\_cell/blob/master/monocle\\_analysis.ipynb](https://github.com/jbloomlab/IFNsorted_flu_single_cell/blob/master/monocle_analysis.ipynb),

2019



Dr Mark Burnley  
@DrMarkBurnley

"I'm the 38th author..."

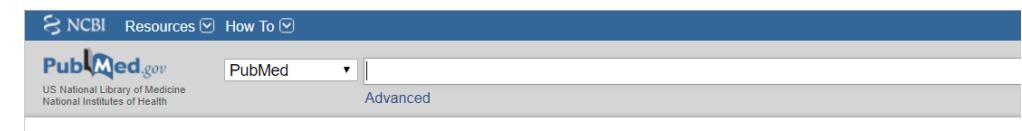
"Wow, that sucks."

"I hadn't finished. I'm the 38th author called "Wang"."

"Oh."

[ncbi.nlm.nih.gov/pubmed/26799652](https://ncbi.nlm.nih.gov/pubmed/26799652)

10:47 PM · Feb 8, 2016 · Twitter Web Client



A screenshot of a Twitter post from Dr. Mark Burnley (@DrMarkBurnley). The post contains four short text snippets and a link to a PubMed article. The link is [ncbi.nlm.nih.gov/pubmed/26799652](https://ncbi.nlm.nih.gov/pubmed/26799652). The timestamp is 10:47 PM · Feb 8, 2016 · Twitter Web Client.

Format Abstract ▾  
Autophagy, 2016;12(1):1-222. doi: 10.1080/15548627.2015.1100356.  
**Guidelines for the use and interpretation of assays for monitoring autophagy (3rd edition).**  
Klionsky DJ, Abdelmohsen K, Abe A, Abedin MJ, Abieliovich H, Acevedo-Arozena A, Adachi H, Adams CM, Adams PD, Adeli K, Adhiketty PJ, Adler SG, Agam G, Agarwal R, Agha MK, Agnello M, Agostinis P, Aguilar PV, Aguirre-Ghiso J, Airoldi EM, Ait-Si-Ali S, Akematsu T, Akporiaye ET, Al-Rubeai M, Albacete GM, Albanese C, Albani D, Albert ML, Aldudo J, Alguí H, Alirezai M, Alloza I, Almasan A, Almonte-Becerril M, Almenr E, Alonso C, Altan-Bonnet N, Altschuler DC, Alvarez S, Alvarez-Erviti L, Alves S, Amadoro G, Amano A, Amantini C, Ambrosio S, Amelio I, Amer AO, Amessou M, Amon A, An Z, Ananias FA, Andersen SU, Andley UP, Andreadi CK, Andreu-Abadie N, Anel A, Ann DK, Anopukumar-Dukie S, Antonioli M, Aoki H, Apostolova N, Aquila S, Aquilano K, Araki K, Arama E, Aranda A, Araya J, Arcaro A, Arias E, Arimoto H, Ariosa AR, Armstrong JL, Arnould T, Arsov I, Asanuma K, Askaras V, Asselin E, Atarashi R, Atherton SS, Atkin JD, Attardi LD, Auburger P, Auburger G, Aurelian L, Autelli R, Avagliano L, Avantaggiati ML, Avrahami L, Awale S, Azad N, Bachetti T, Backer JM, Bae DH, Bae JS, Bae ON, Bae SH, Baehrecke EH, Baek SH, Baghdiguian S, Baghresa-Zadworna A, Bai H, Bai J, Bai XY, Bally Y, Balaji KN, Balduini W, Ballabio A, Balzan R, Banerjee R, Bánhegyi G, Bao H, Barbeau B, Barrachina MD, Barreiro E, Bartel B, Bartolomé A, Bassham DC, Bassi MT, Bast RC Jr, Basu A, Batista MT, Batoko H, Battino M, Bauckman K, Baumgartner BL, Bayer KU, Beale R, Beaulieu JF, Beck GR Jr, Becker C, Beckham JD, Bédard PA, Bednarski PJ, Begley TJ, Behl C, Behrends C, Behrens GM, Behrns KE, Bejarano E, Belaid A, Belleudi F, Bénard G, Berchem G, Bergamaschi D, Bergami M, Berkhouit B, Berliocchi L, Bernard A, Bernard M, Bernassola E, Bertolotti A, Bess AS, Besteiro S, Bettuzzi S, Bhalla S, Bhattacharyya S, Bhutia SK, Biagiosch C, Bianchi MW, Biard-Piechaczyk M, Billes V, Bincoletto C, Bingol B, Bird SW, Bitoun M, Bjedov I, Blackstone C, Blanc L, Blanco GA, Blomhoff HK, Boada-Romero E, Böcker S, Boes M, Boesze-Battaglia K, Boisse LH, Bolino A, Roman A, Bonaldo P, Bordi M, Bosch J, Botana LM, Bott J, Bou G, Bouché M, Bouchebarell M, Boucher MJ, Boulton ME, Bouret SG, Boyer P, Boyer-Guittaut M, Bozhkov PV, Brady N, Braga VM, Brancolini C, Braus GH, Bravo-San Pedro JM, Brennan LA, Bresnick EH, Brest P, Bridges D, Bringer MA, Brini M, Brito GC, Brodin B, Brookes PS, Brown EJ, Brown K, Broxmeyer HE, Bruhat A, Brum PC, Brumell JH, Brunetti-Pierri N, Bryson-Richardson RJ, Bush C, Butuk H, Buvatin DV, Bultman SJ, Bulyntck G, Bumbasirevic V, Burelle Y, Burke RE, Burmeister M, Butikofer P, Caberlotto L, Cadwell K, Cahoy M, Cai D, Cai J, Cai Q, Calatayud S, Camougrand N, Campanella M, Campbell GR, Campbell M, Campello S, Candau R, Caniggi J, Cantoni L, Cao L, Caplan AB, Caraglia M, Cardinali C, Cardoso SM, Carew JS, Carlton LA, Carlton CR, Carloni S, Carlsson SR, Carmona-Gutierrez D, Carneiro LA, Carnevali O, Carras A, Carrier A, Carroll B, Casas C, Casas J, Cassinelli G, Castets P, Castro-Obregon S, Cavallini G, Ceccherini I, Ceconi F, Cederbaum AI, Cefnia V, Censi S, Cerella C, Cervia D, Cetrullo S, Chaachouay H, Chae HJ, Chagin AS, Chai CY, Chakrabarti G, Chamilos G, Chan EY, Chan MT, Chandra D, Chandra P, Chano RC, Chang TY, Chatham JC, Chatterjee S, Chauhan S, Che

jbloom saved plot of association between co-infection / IFN

1 contributor

12.8 MB

## Table of Contents

- [Analyze viral features associated with IFN induction](#)
  - [Setup for analysis](#)
    - [Load / install packages](#)
    - [Notebook-wide variables / functions](#)
  - [Get cell-gene matrices](#)
    - [Specify cell types](#)
    - [Load cell-gene matrix](#)
  - [Count cells and annotate multiplets](#)
    - [Annotate cross-celltype multiplets](#)
    - [Number of cells and multiplet frequency](#)
    - [Plot summarizing cell counts and multiplets](#)
  - [Filter multiplets and low-quality cells](#)
    - [Remove cross-celltype multiplets](#)
    - [Number of cellular and flu mRNAs, bounds for filtering](#)
    - [Plot cellular / flu mRNAs with filters](#)
    - [Filter cells with extreme mRNA amounts](#)
  - [Call infection / gene presence from canine cell thresholds](#)
    - [Constant fraction or number of mRNAs from flu?](#)
    - [Confirm equal mix of flu barcodes in canine cells](#)
    - [Look at segment frequencies](#)
    - [Get human cells for infection-status calling](#)
    - [Compute P-value flu is above background](#)
    - [Call infected cells by amount of total flu](#)
    - [Call gene presence/absence](#)

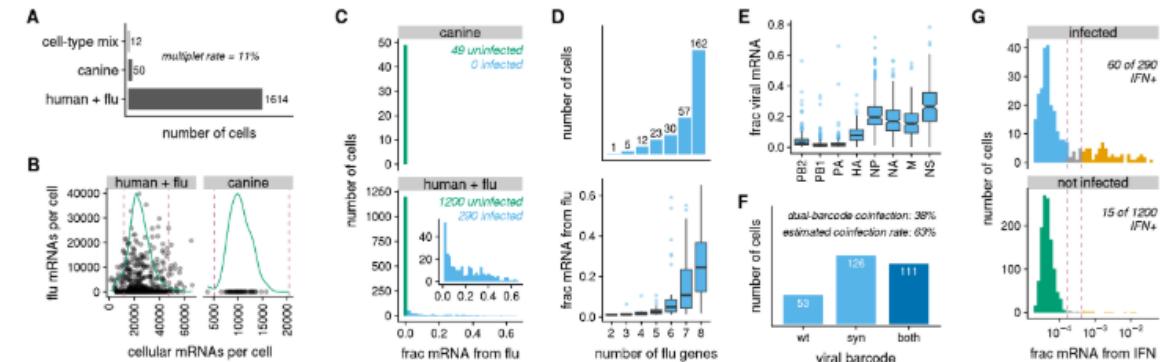
## Figures for paper

We have made all of the plots above, and saved some of them to the figures directory already by using the `isfig=TRUE` argument to `saveShowPlot`. However, there are others that we want to assemble into multi-panel figures. We do that here.

First, we assemble a figure that shows the calling of cells, infected cells, and IFN+ cells:

```
In [101]: p_cell_summary <- plot_grid(
  plot_grid(p_cellcounts, p_fiu_vs_cell,
            ncol=1, rel_heights=c(1, 1.5), scale=0.9,
            labels=c("A", "B"), label_size=18, vjust=1),
  plot_grid(p_frac_fiu, labels="C", scale=0.95, label_size=18, vjust=1),
  plot_grid(p_nfui_genes, labels="D", scale=0.95, label_size=18, vjust=1),
  plot_grid(p_fiu_rel_expr, p_cinfect,
            scale=0.95, ncol=1,
            labels=c("E", "F"), label_size=18, vjust=1),
  plot_grid(p_ifn_dist, labels="G", scale=0.95, label_size=18, vjust=1),
  nrow=1, scale=0.95, rel_widths=c(1, 0.7, 0.6, 0.75, 0.7), align="h"
) +
  theme(plot.margin=unit(c(t=0, r=0, b=-0.3, l=0), "in"))

saveShowPlot(p_cell_summary, width=15.5, height=4.9, isfig=TRUE)
```



Now a supplementary figure to the one above with the single-cell transcriptomic data:

```
In [102]: p_cell_summary_supp <- plot_grid(
  p_frac_has_gene,
  p_ifn_genes_corr,
  p_isg_dist,
  p_isg_corr,
  ncol=2,
  scale=0.9,
  rel_heights=c(0.68, 1),
  labels=c("A", "B", "C", "D"), label_size=18, vjust=2, hjust=-1
)

saveShowPlot(p_cell_summary_supp, width=9.5, height=7.5)
```



# Putting the final pieces into place

## Make Your Code Citable Using GitHub and Zenodo: A How- to Guide

By [Open Science MOOC](#) on July 24, 2018



# Acknowledgements/Local Supports



**Dr Darren Dahly**

**Dr Brendan Palmer**

The image is a screenshot of the University College Cork (UCC) Library Research Data Service website. At the top, there's a dark blue header bar with white text for 'Home', 'Subject Support', 'Services', 'Ask Us', and 'My Account'. Below the header, the page title 'Research Data Service: Home' is displayed. A sub-header below it states: 'We are a university wide resource which supports and promotes best practice in data management.' Underneath, there's a navigation menu with tabs for 'Home' (which is highlighted in blue), 'Data Management Planning', 'FAIR', 'Training and Support', and 'Open Science'. A decorative graphic of a winding path leads to the text 'Research Data Service'. The overall design is clean and professional, using a light grey background for the main content area.

**Eoghan Ó Carragáin**

**Dr Aoife Coffey**