Data FAIRification using RStudio workflows

Brendan Palmer & Darren Dahly,
Clinical Research Facility - Cork & School of Public Health,
University College Cork
@B_A_Palmer @statsepi

CRF-C
HRB Clinical Research Facility Cork

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh
School of Public Health

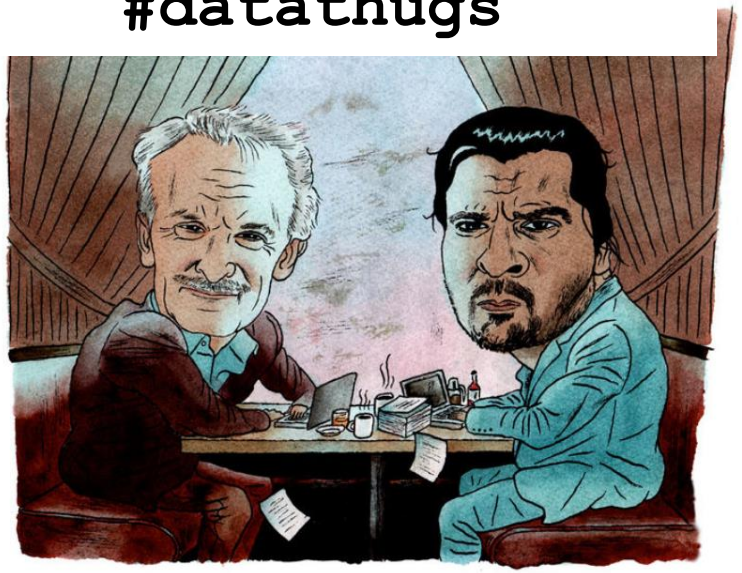
Where to begin…

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

E-alert RSS Facebook Twitter

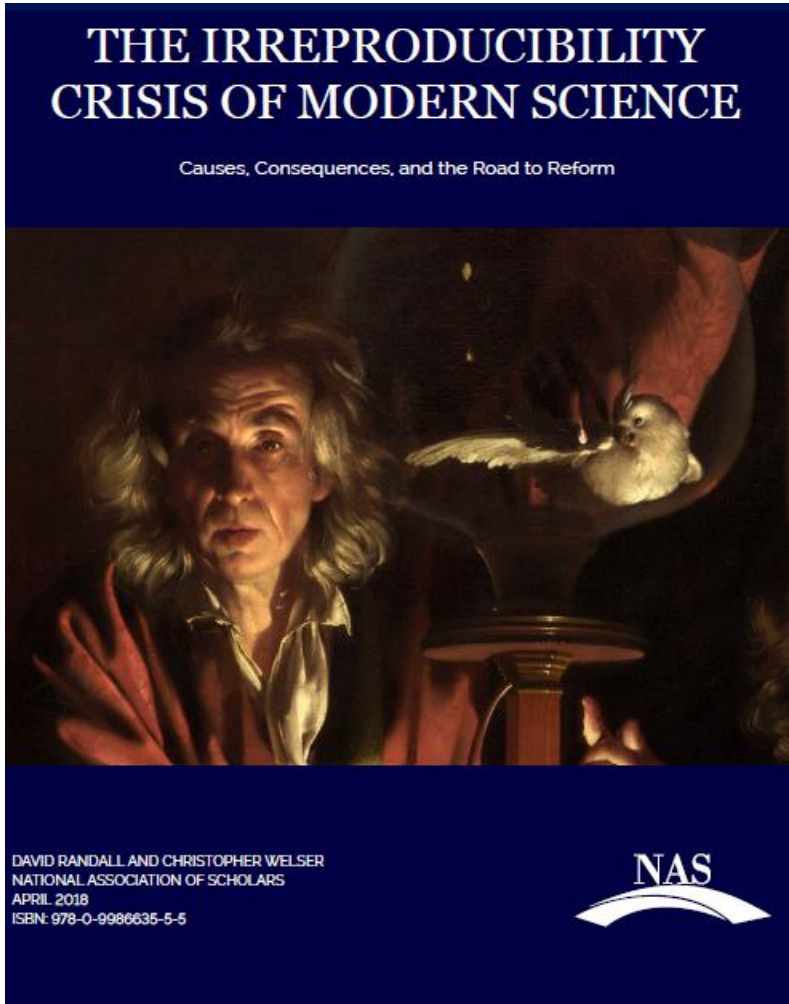
NATURE | NEWS FEATURE

1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

Monya Baker

2018



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

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 520 > Issue 7549 > Comment > Article

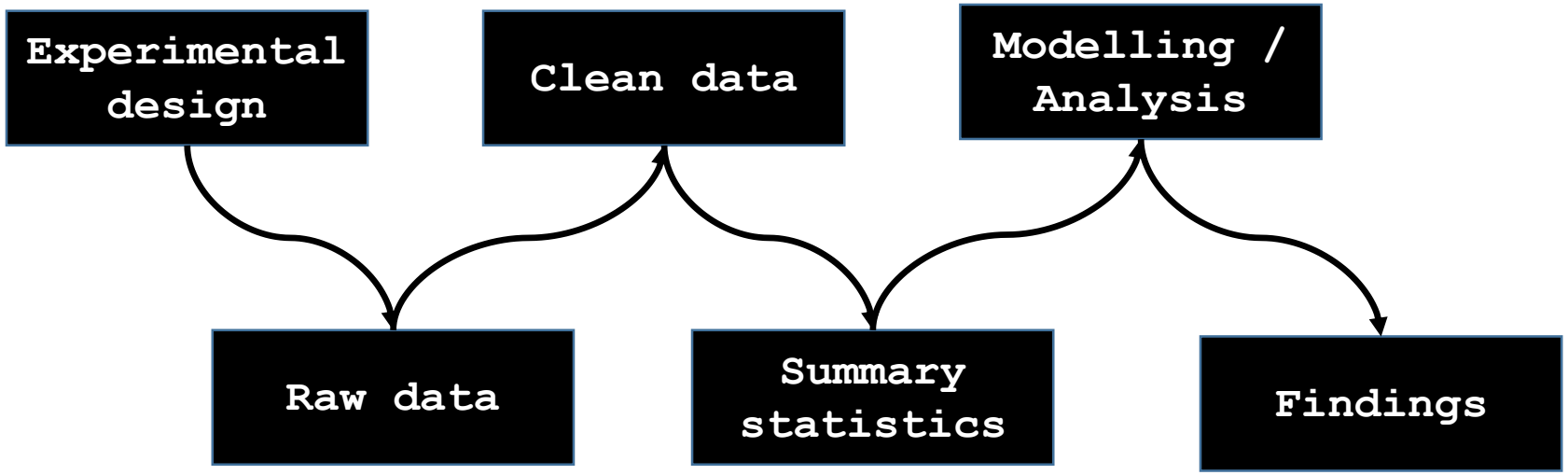
E-alert RSS Facebook Twitter

NATURE | COMMENT

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

Jeffrey T. Leek & Roger D. Peng

28 April 2015

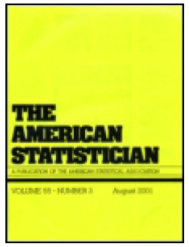


The winds of change

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.



Taylor & Francis

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

Taylor & Francis

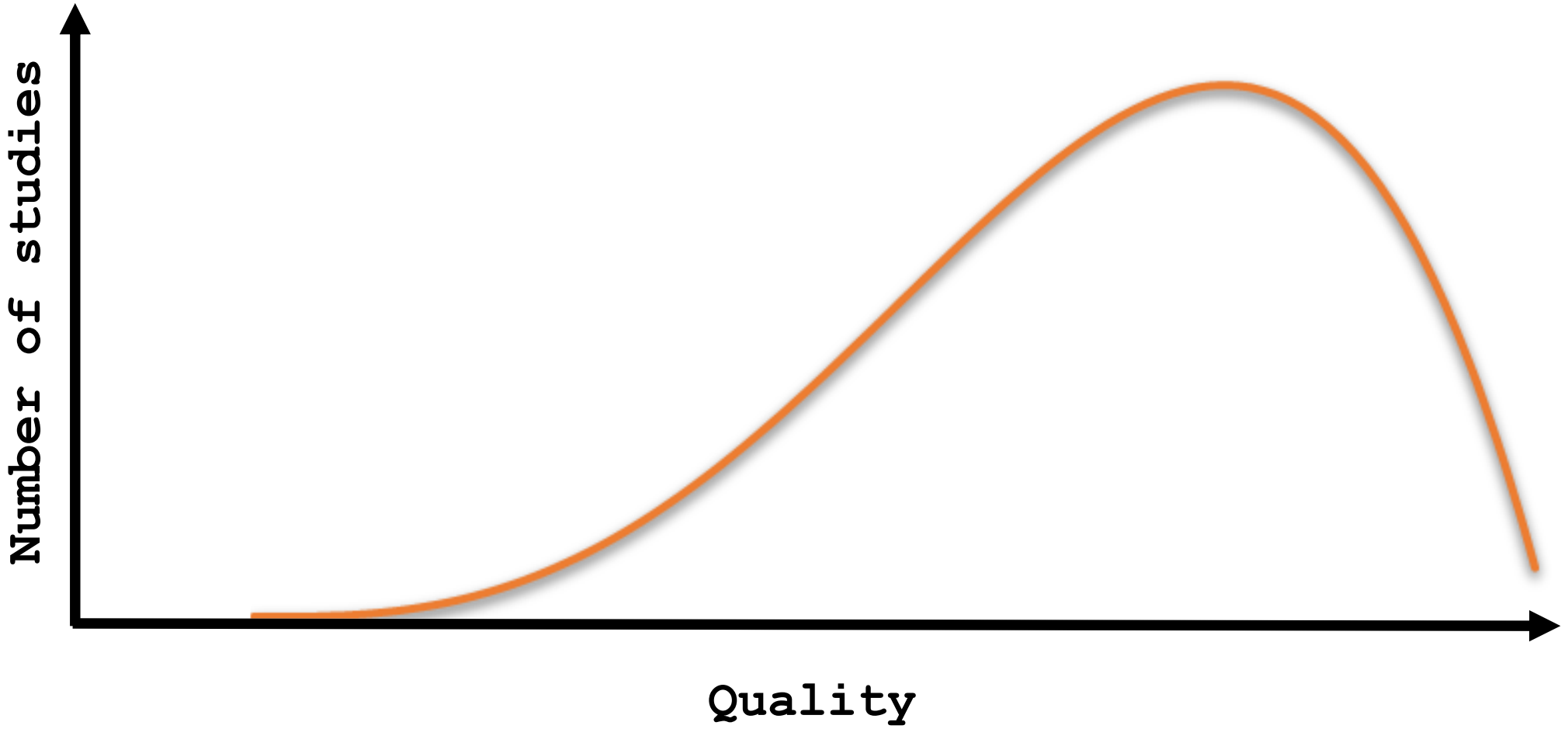
The American Statistician

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

Plan S
Making full and immediate Open Access a reality

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."

HRB Health Research Board

Funding | Data collections & evidence | Publications | Success stories | News | About

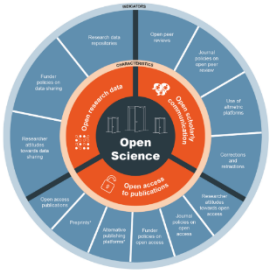
Home > Funding > Policies and principles > Open Research

- Funding schemes
- EU funding support
- Manage a grant
- Funding awarded
- Evaluation
- GDPR guidance for researchers
- Policies and principles
  - EU legislation
  - Gender
  - Good research practice
  - Open Research

Open Research

The HRB is committed to ensuring that its funded research is open, accessible and usable, so it can have the greatest possible impact.

There is a fundamental shift across Europe towards making research more transparent, collaborative, accessible and efficient. This Open Science movement is a strategic priority for the European Commission in research and innovation policy and an EU high-level Expert Group, the Open Science Policy Platform (OSPP 2016-2018) has been established to consider key implementation areas.



EUROPEAN COMMISSION
Directorate-General for Research & Innovation

H2020 Programme

Guidelines on

FAIR Data Management in Horizon 2020

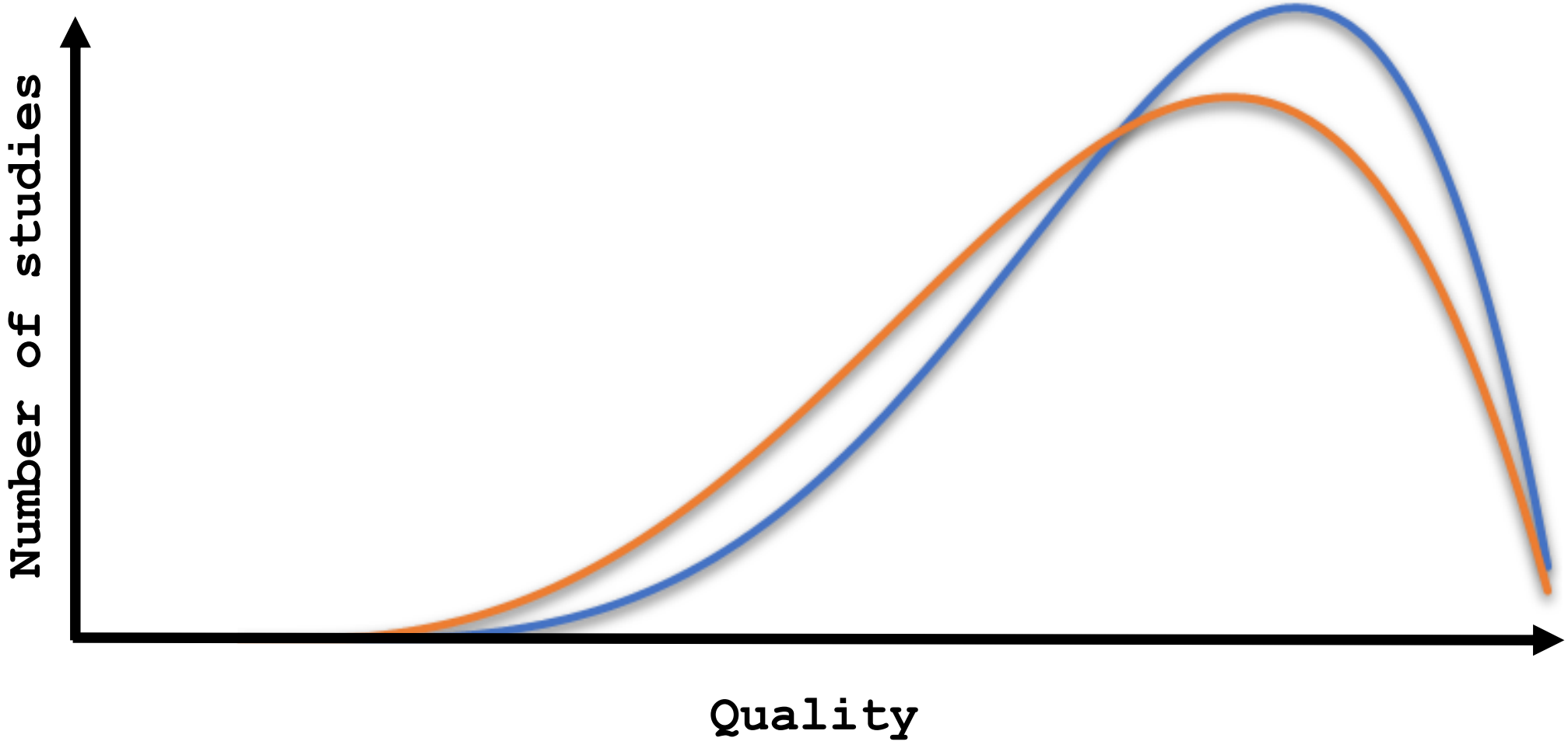
Science Foundation Ireland sfi For what's next

Funding | Engagement | Events | Research News | SFI Research Centres

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.

Tomorrow



FAIR is a part of your life now!

Data Guidelines

wellcome

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

Science Foundation Ireland sfi For what's next

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.

HRB Health Research Board

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**.

UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh

Social Research Ethics Committee (SREC)

ETHICS APPROVAL FORM

srec@ucc.ie
https://www.ucc.ie/en/research/about/ethics/

NIH National Institutes of Health
Turning Discovery Into Health

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).

[4] 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.

But what does that mean?

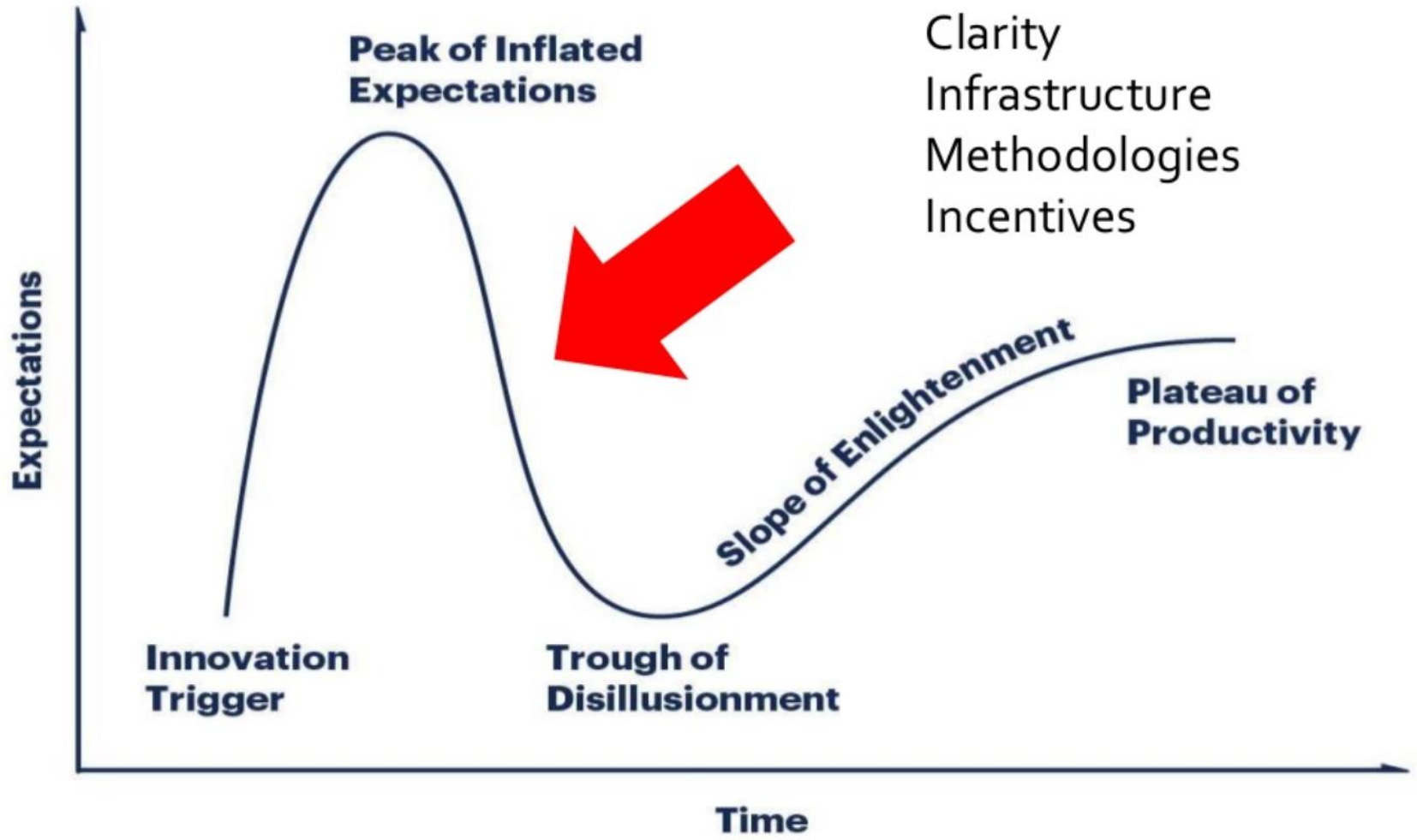
Naomi Penfold
@npscience

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

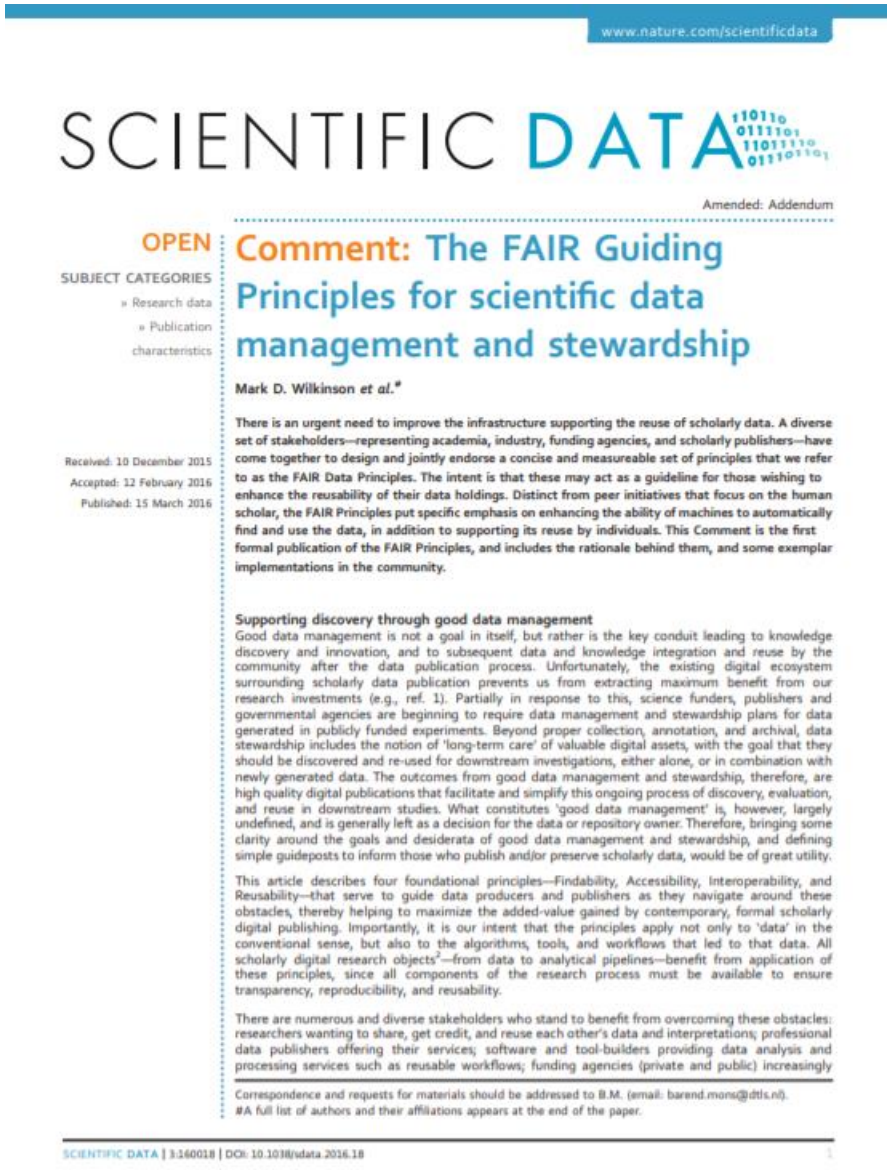
12:57 PM · Sep 1, 2019 · Twitter Web App

FAIR Safari – Prof Carole Goble

The FAIR Hype



What are the FAIR data principles



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 measureable 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[2]—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

Findable - Assign persistent IDs
- Machine readable descriptions to support structured searches

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

Interoperable - Uses standard (FAIR) vocabularies
- Linked to other resources

Reusable - Clear licenses
- Provenance
- Meets domain-relevant community standards

A path towards FAIR

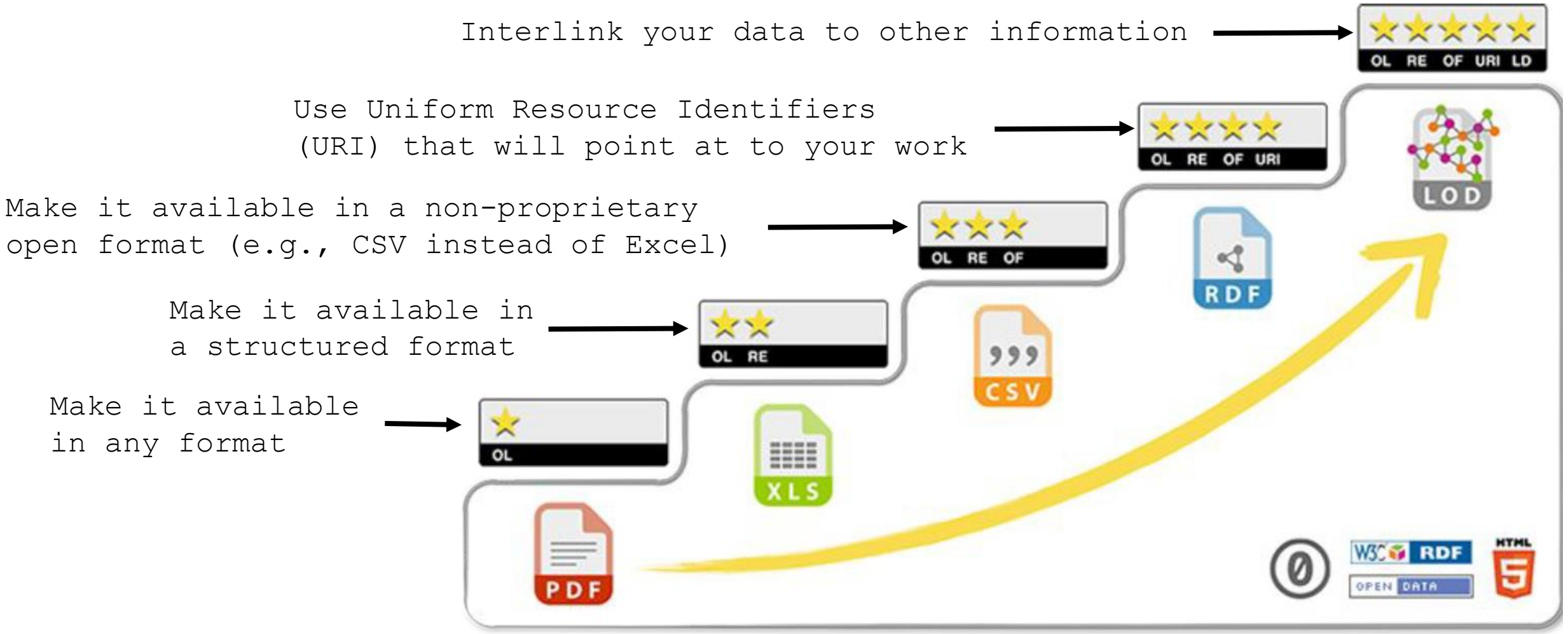


Image: 5 ★ OPEN DATA

Advanced FAIRification – linked vocabularies

VOCABS TERMS AGENTS SPARQL/DUMP

Linked Open Vocabularies (LOV)

+ Suggest | Documentation | Follow

DCMI Metadata Terms (dcterms)

Metadata

URI	http://purl.org/dc/terms/
Namespace	http://purl.org/dc/terms/
homepage	http://dublincore.org/documents/dcmi-terms
Description	an up-to-date specification of all metadata terms maintained by the Dublin Core Metadata Initiative, including properties, vocabulary encoding schemes, syntax encoding schemes, and classes. @en
Language	English en
Creator	Dublin Core Metadata Initiative http://purl.org/dc/aboutdcmi#DCMI
Publisher	Dublin Core Metadata Initiative http://purl.org/dc/aboutdcmi#DCMI
Comment	(2013-03-07) Bernard Vatant: Prefix restored to dcterms (2014-03-14) Bernard Vatant: This vocabulary is one of the most used in the LOD cloud, and here to stay, even if the purl redirection is sometimes down, like at the time I write this review. (2015-03-24) Bernard Vatant: Annual review OK (2016-05-10) Ghislain Atemezing: Annual review OK (2018-08-02) Ghislain Atemezing: Annual review - OK

n3

Statistics

Classes 34
Properties 55
Datatypes 12
Instances 1

Expressivity

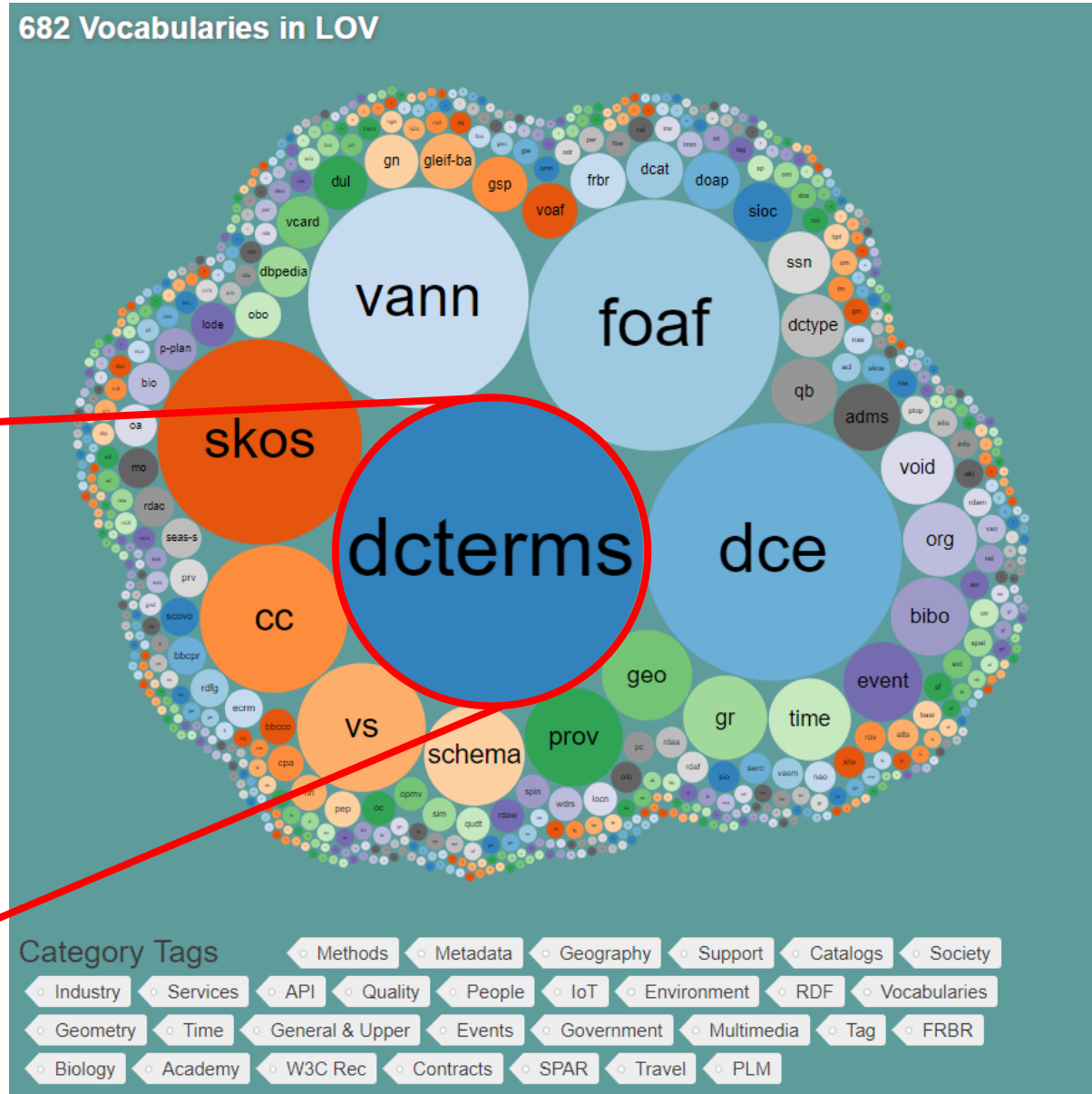
RDF RDFS OWL

Tags

Metadata

LOD

Vocabulary used in **327 datasets**



RDF Data

By input | By URL | By File | By Endpoint

```
@prefix schema: <http://schema.org/> .
@prefix :       <http://example.org/> .
@prefix xsd:    <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf:   <http://xmlns.com/foaf/0.1/> .

:alice  schema:gender   schema:Female ;
        schema:knows    :bob ;
        schema:name     "Alice" .

:bob    schema:birthDate  "1980-03-10"^^xsd:date ;
        schema:gender     schema:Male ;
        schema:name       "Robert" .

:carol  schema:gender   schema:Female ;
        schema:name     "Carol" ;
        foaf:age        23 .
```

Incorporating FAIR into your routine workflow

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

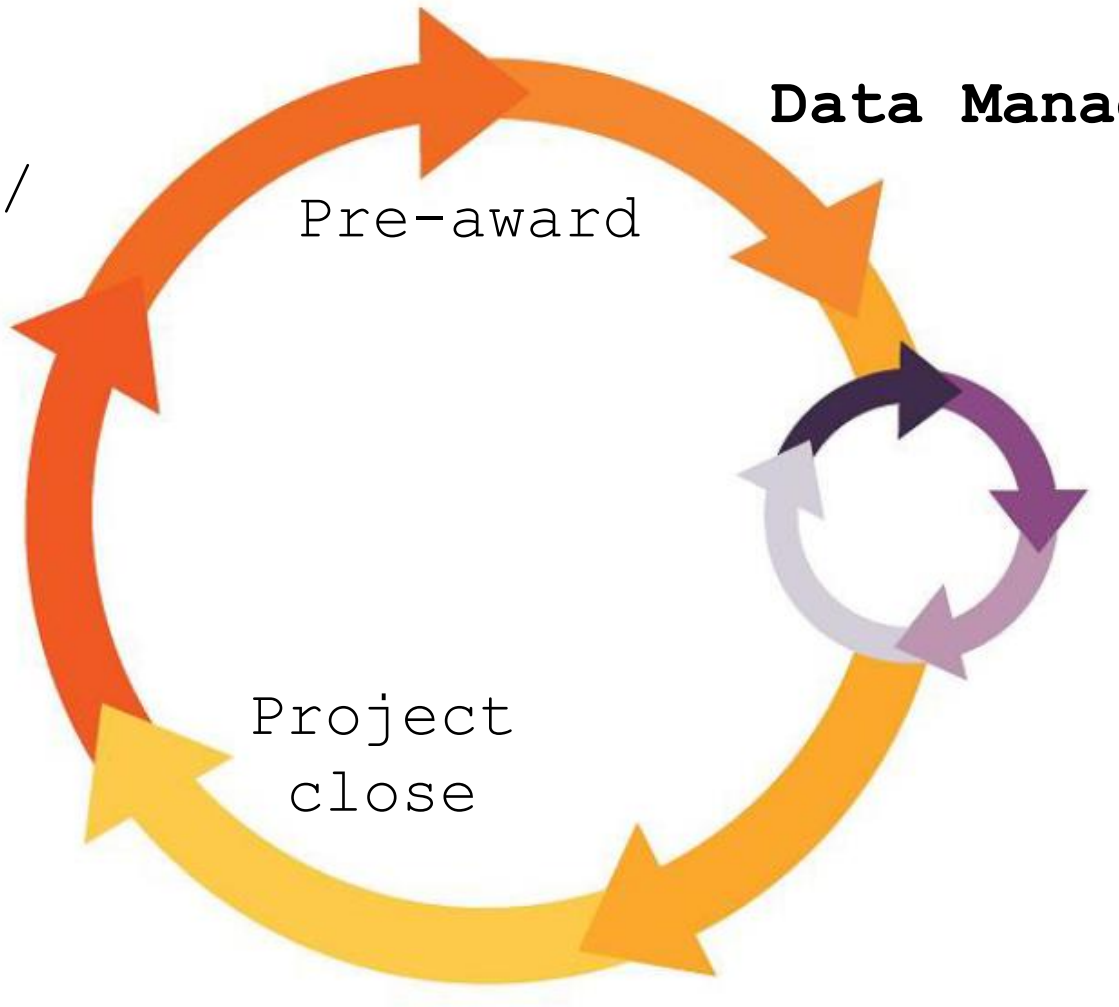
An output management plan (OMP) is a useful starting point for collecting or creating data, software, research materials, and intellectual property. Creating an OMP before you begin your research, and updating it throughout the research cycle, will help ensure that your outputs are as open and **FAIR**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

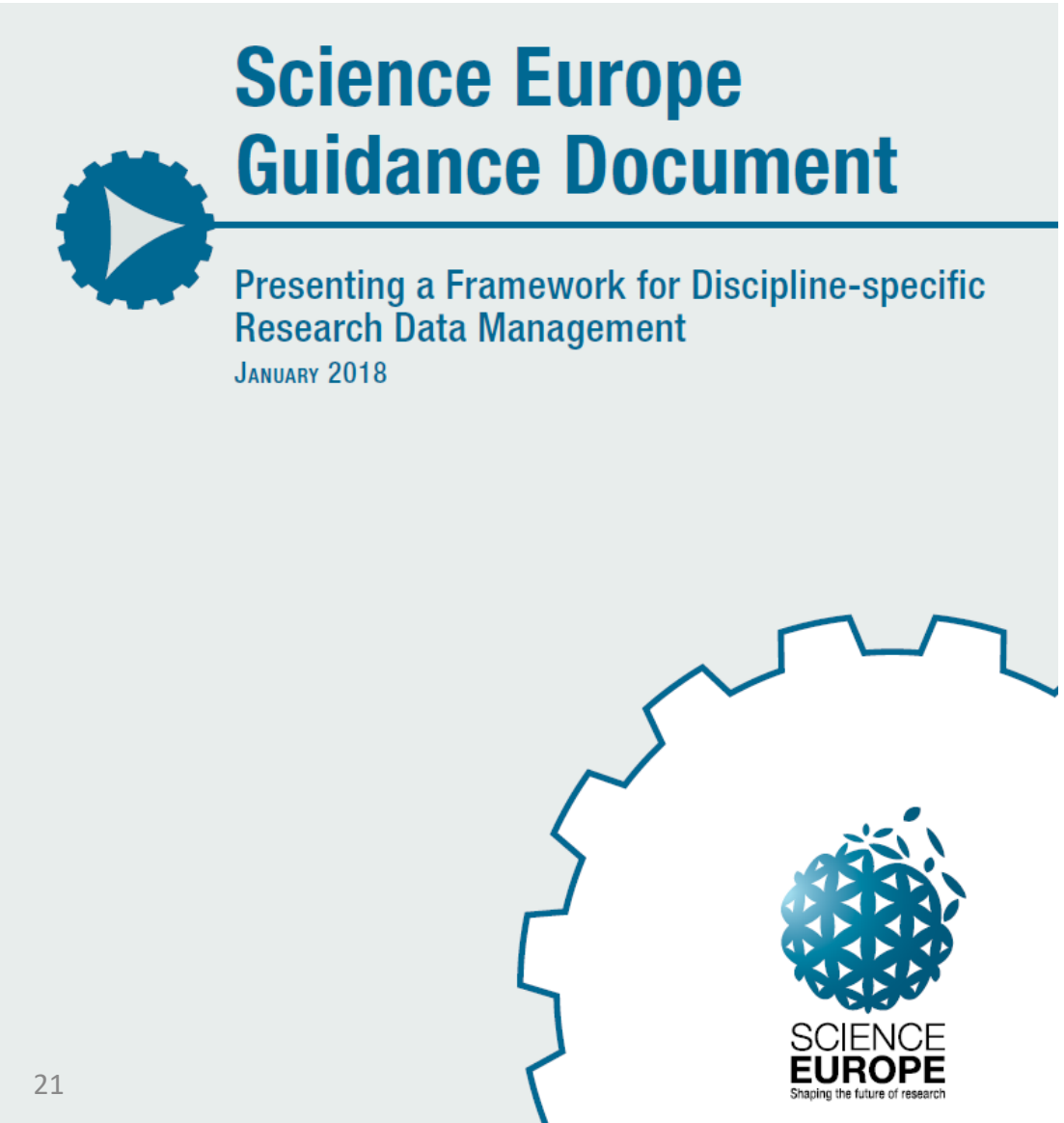
You should consider:

- 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



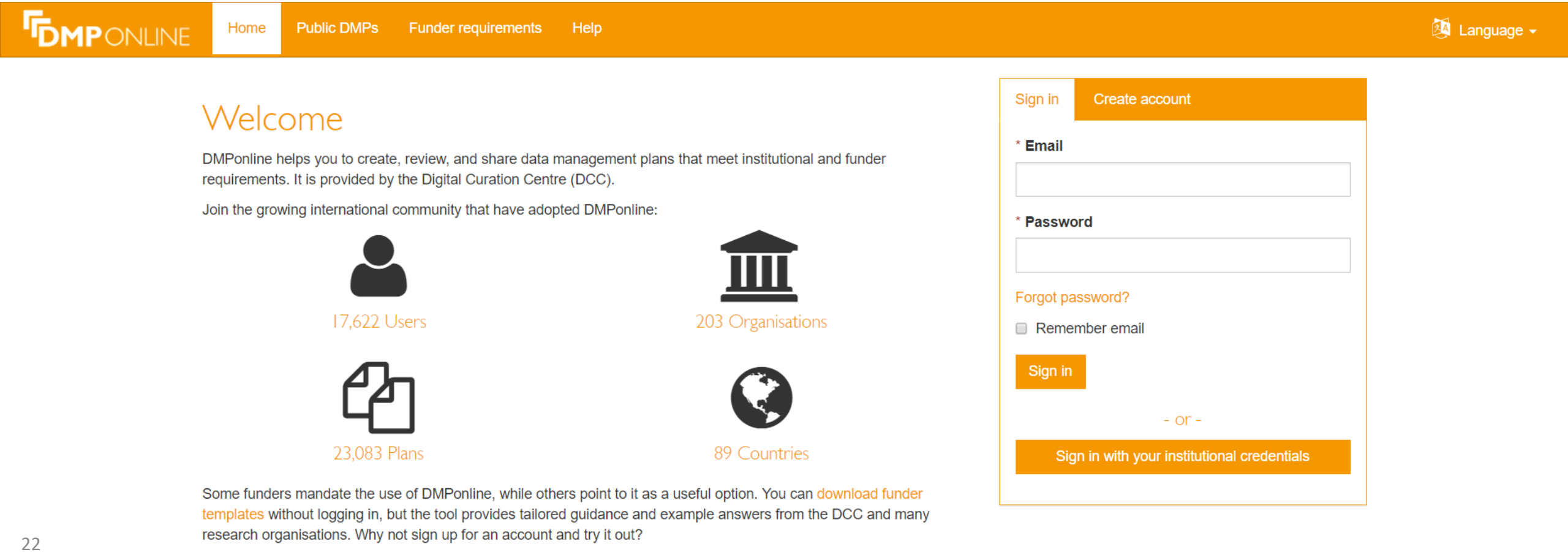
It all starts with a Data Management Plan



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



DMP ONLINE | My Dashboard | Create plans | Reference | Help | Language | Brendan Palmer

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

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
For exploring the DSW features	*For individual researchers*	*For organizations*	*For organizations*
• Easy to sign up and use • A shared instance with other users • **Not for serious usage**	• Easy to sign up and use • Ready to use Knowledge Models • Privacy and stability	• All the DSW features available • Your own instance • You need to host and run the instance by yourself	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.

Exploring standards in your field

FAIRsharing.org
standards, databases, policies

Search all of FAIRsharing | Standards | Databases | Policies | Collections | Add/Claim Content | Stats | Log in or Register

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

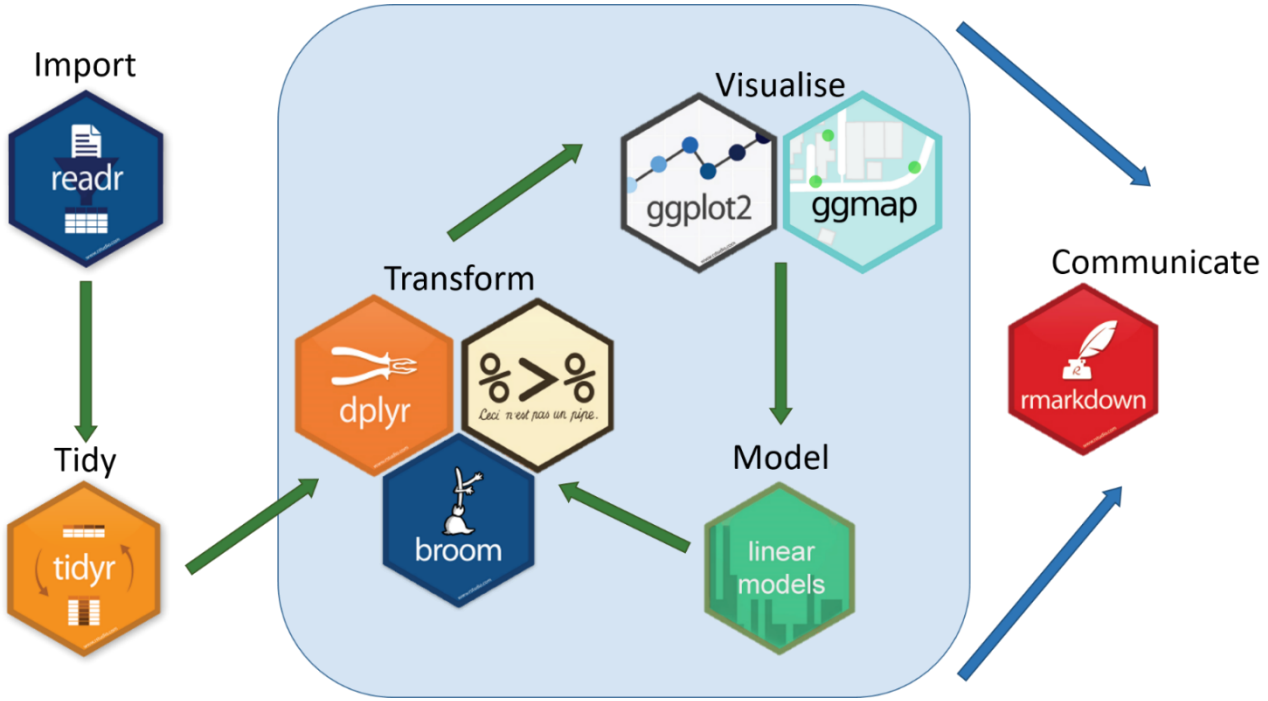
HOW CAN WE HELP?

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.

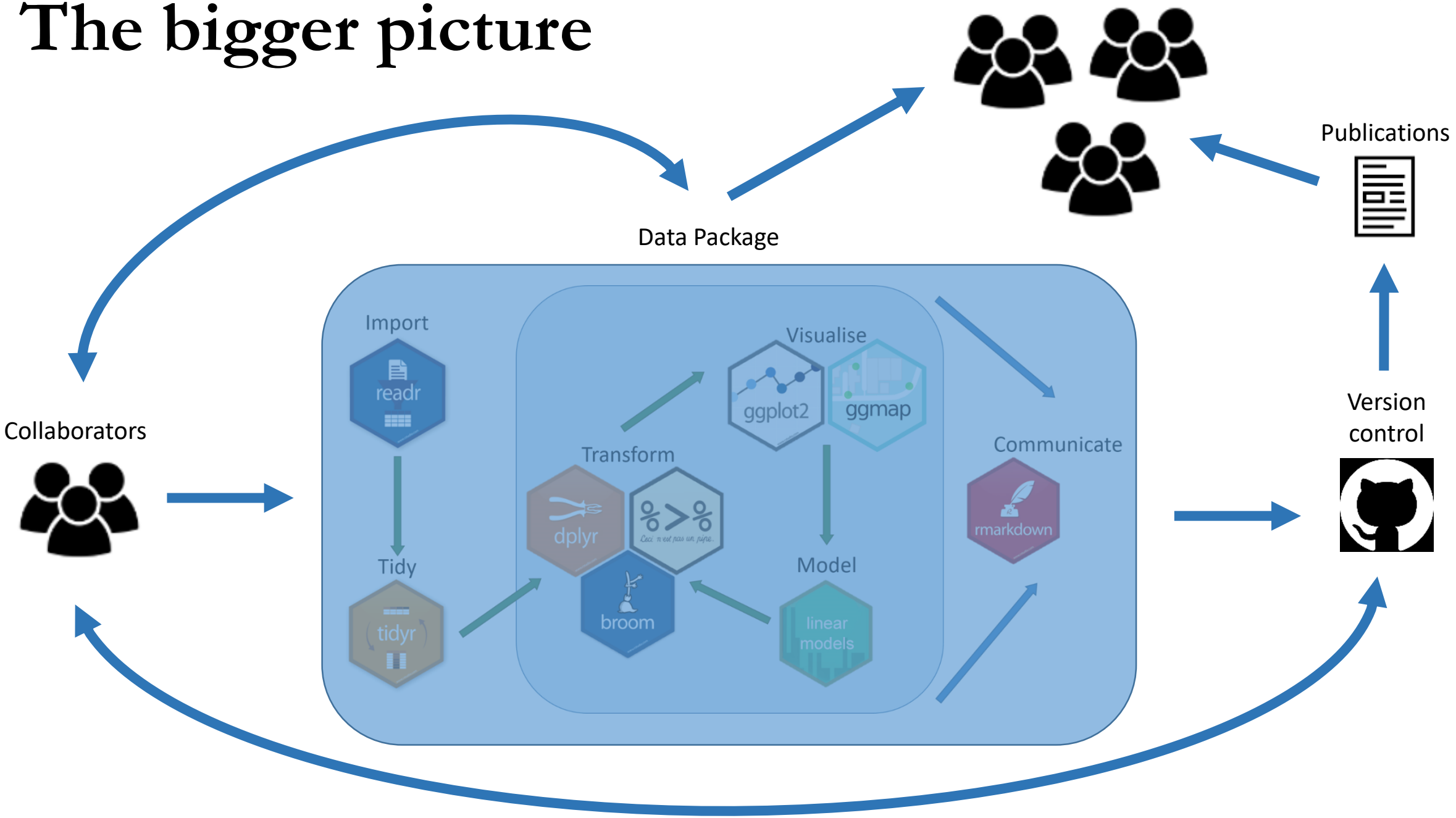
Research data facilitators, librarians, trainers

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]

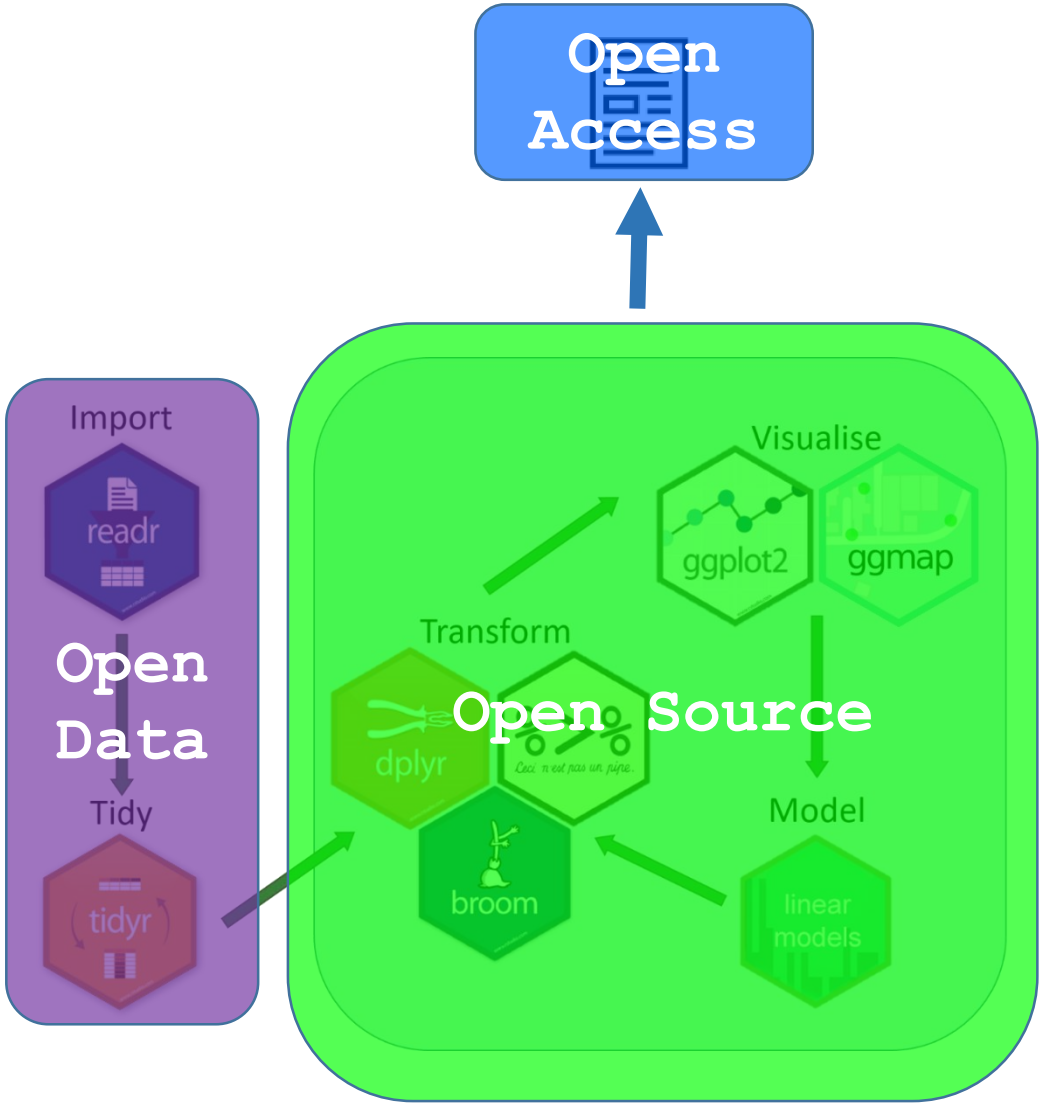
Putting the pieces together using R



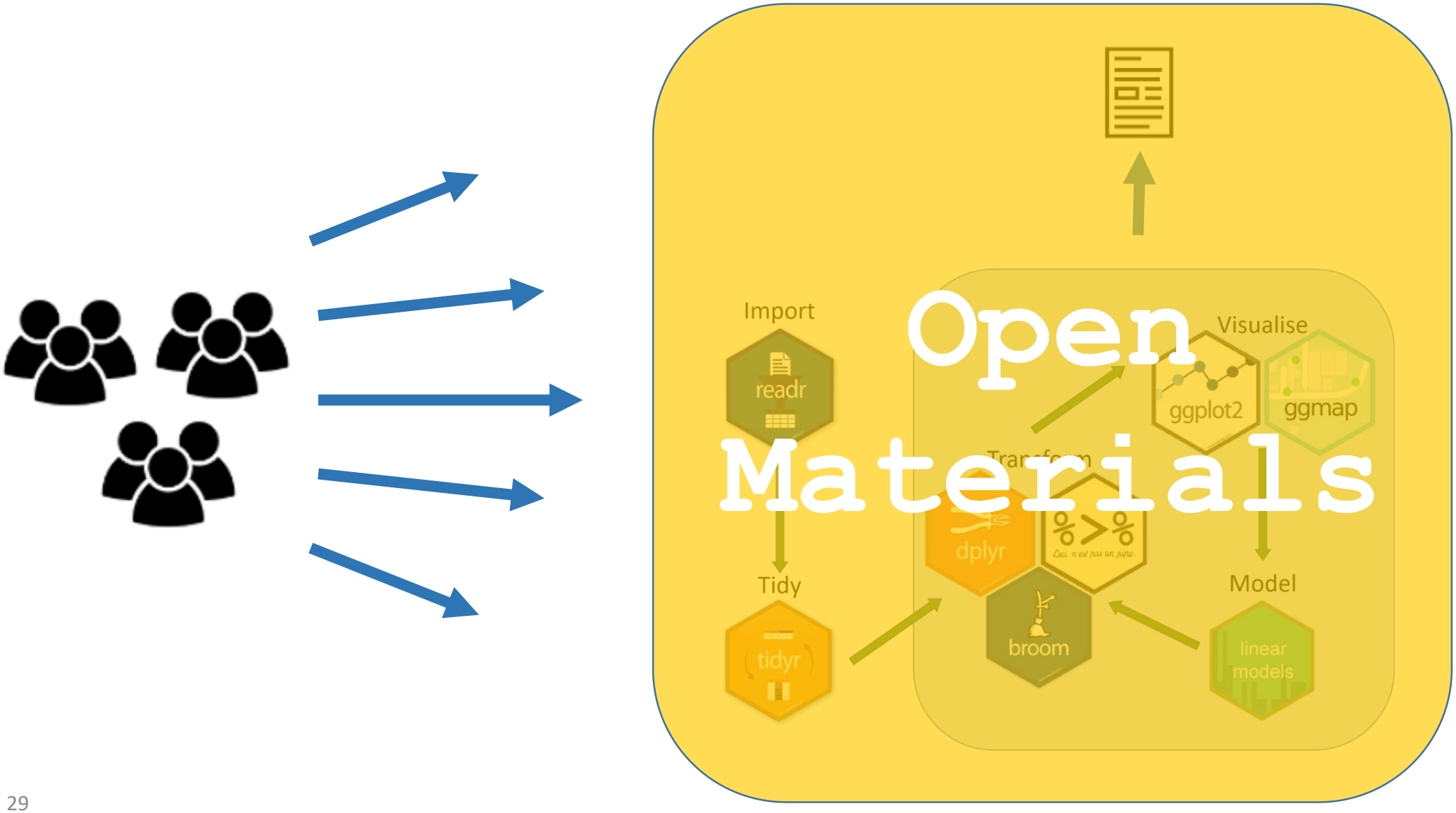
The bigger picture



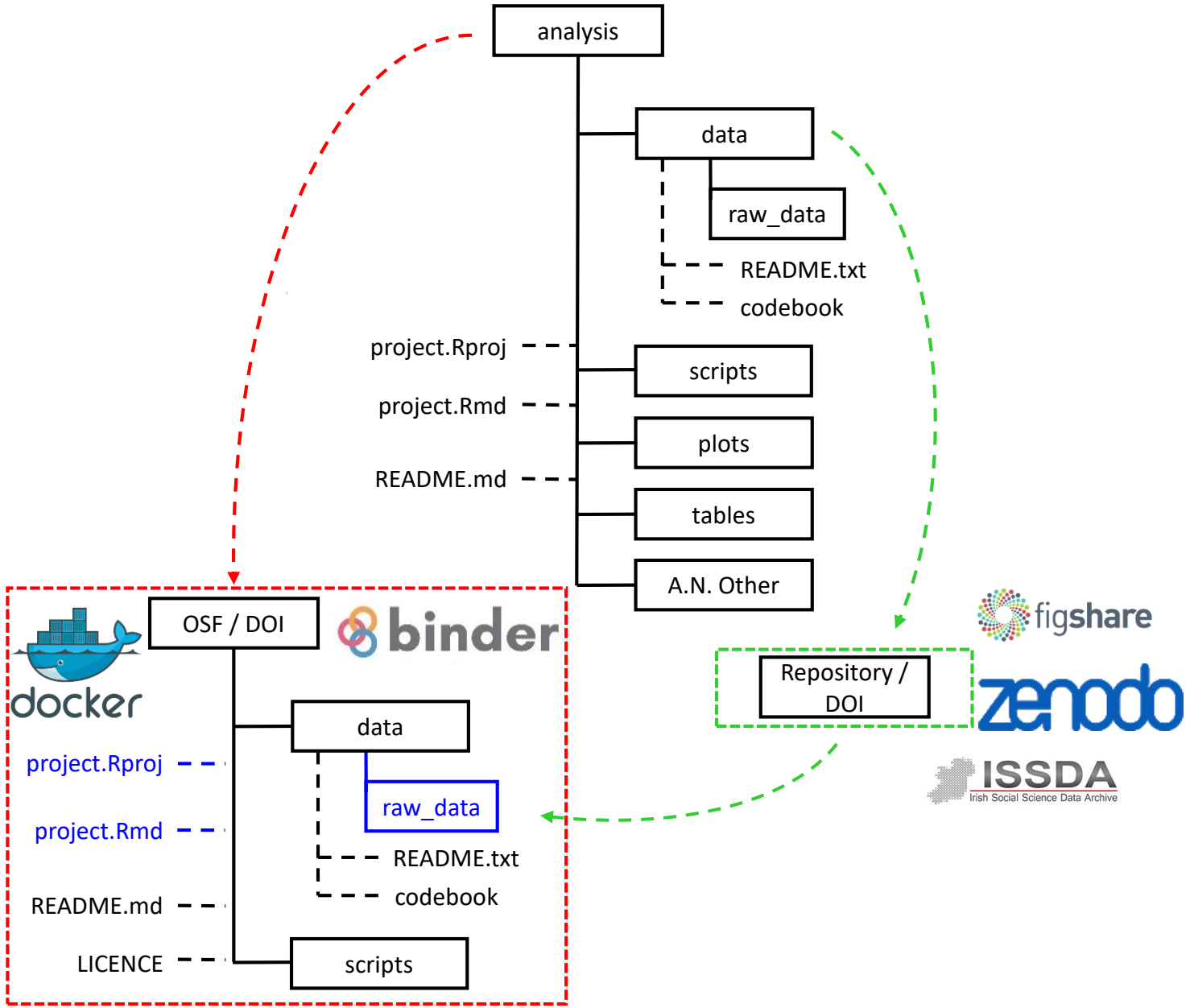
The 'Open Science' picture



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 repository name or URL | GitHub

Git branch, tag, or commit

Git branch, tag, or commit

Path to a notebook file (optional)

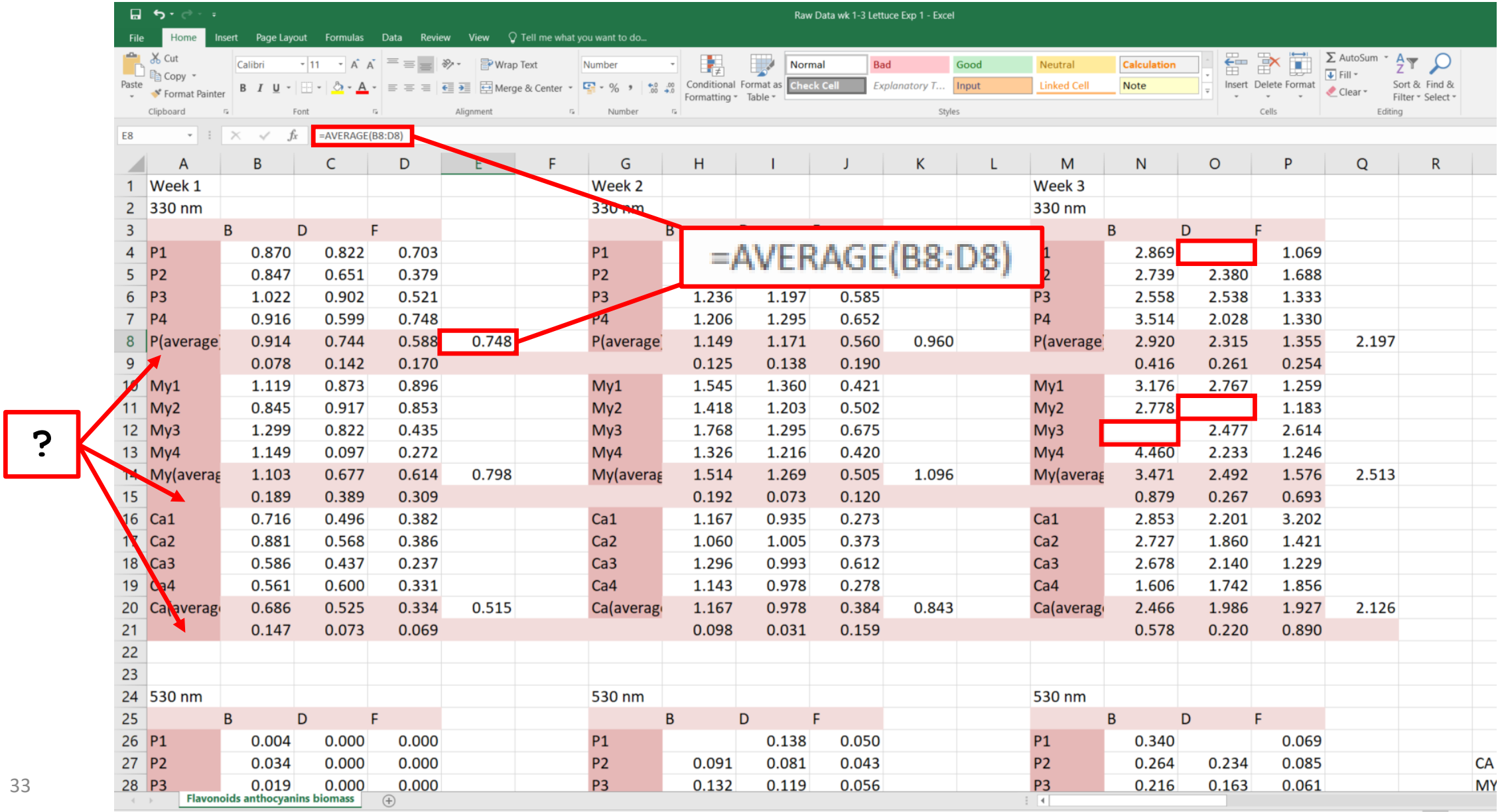
Path to a notebook file (optional) | File | launch

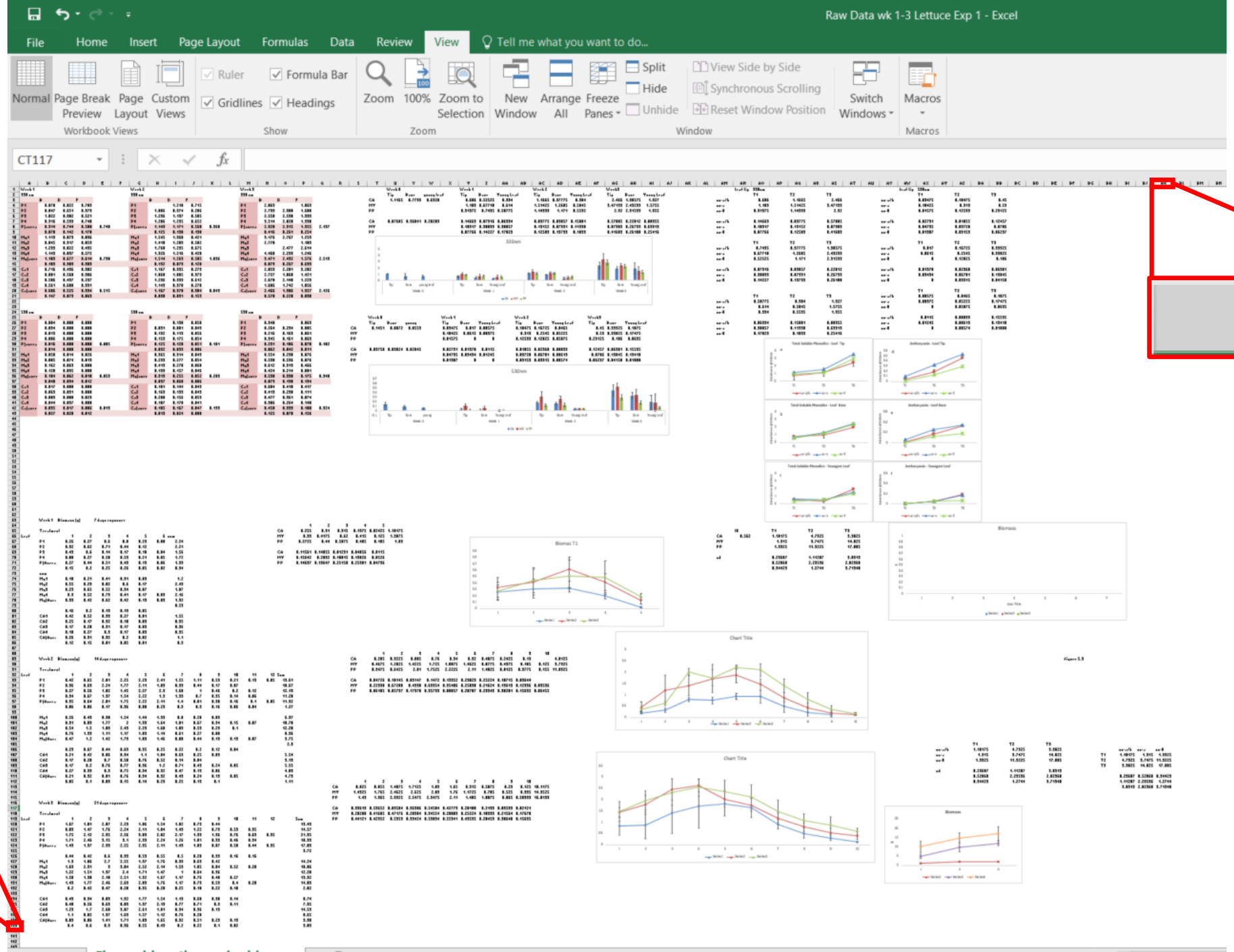
Our real life experiment

- 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
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Tell me what you want to do...
Normal
Page Break Preview
Page Layout
Custom Views
Workbook Views
Ruler
Gridlines
Formula Bar
Headings
Show
Zoom
100%
Zoom to Selection
Zoom
New Window
Arrange All
Freeze Panes
Split
Hide
Unhide
View Side by Side
Synchronous Scrolling
Reset Window Position
Window
Switch Windows
Macros
Macros
CT117
BK
139
Flavonoids anthocyanins biomass

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

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]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

	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	my	filter	1	1.545	2.45	cos	2019/04/15	Brendan Palmer		

data | dictionary | values

Less stress, more success

id	week_no	filter_nam	treatment	replicate_no	flavonoids	biomass	variety	date	investigator
1	0	ptp	nofilter	1	1.061	0.39	cos	2019/04/01	Darren Dahly
2	0	ptp							
3	0	ptp							
4	0	ptp							
1	0	my							
2	0	my							
3	0	my							
4	0	my							
1	0	ca							
2	0	ca							
3	0	ca							
4	0	ca							
5	1	ptp							
6	1	ptp							
7	1	ptp							
8	1	ptp							
9	1	my							
10	1	my							
11	1	my							
12	1	my							
13	1	ca							
14	1	ca							
15	1	ca							
16	1	ca							
17	2	ptp							
18	2	ptp							
19	2	ptp							
20	2	ptp							
21	2	my							

data | dictionary

field_name	data_type	data_format	example	standard_units	description
id	numeric	integer	23	NA	Unique identifier applied to each observation
week_no	numeric	integer	1	NA	Week number, 1 = 7 days exposure, 2 = 14 days exposure
filter_name	character	NA	my	NA	3 filter types; 'ptp' = polytunnel plastic blocks all UV light
treatment	character	NA	filter	NA	Presence or absence of a filter at the time of sampling
replicate_no	numeric	integer	1	NA	The number of replicates in each treatement
flavonoids	numeric	double	0.3421	parts per million (ppm)	Leaf disc taken from the tip of the most mature leaf at th
biomass	numeric	double		gram (g)	Above ground biomass on the day of harvest
variety	character	NA	cos	NA	3 commerical varieties of red lettuce used; 'cos' = Cos Di
date	date	YYYY/MM/DD	2019/06/28	ISO 8601	Experiment date
investigator	character	Firstname Lastname	Aoife Coffey	NA	Primary researcher who performed the experiment

data | dictionary | values

Less stress, more success

	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									
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	my									

data | dictionary

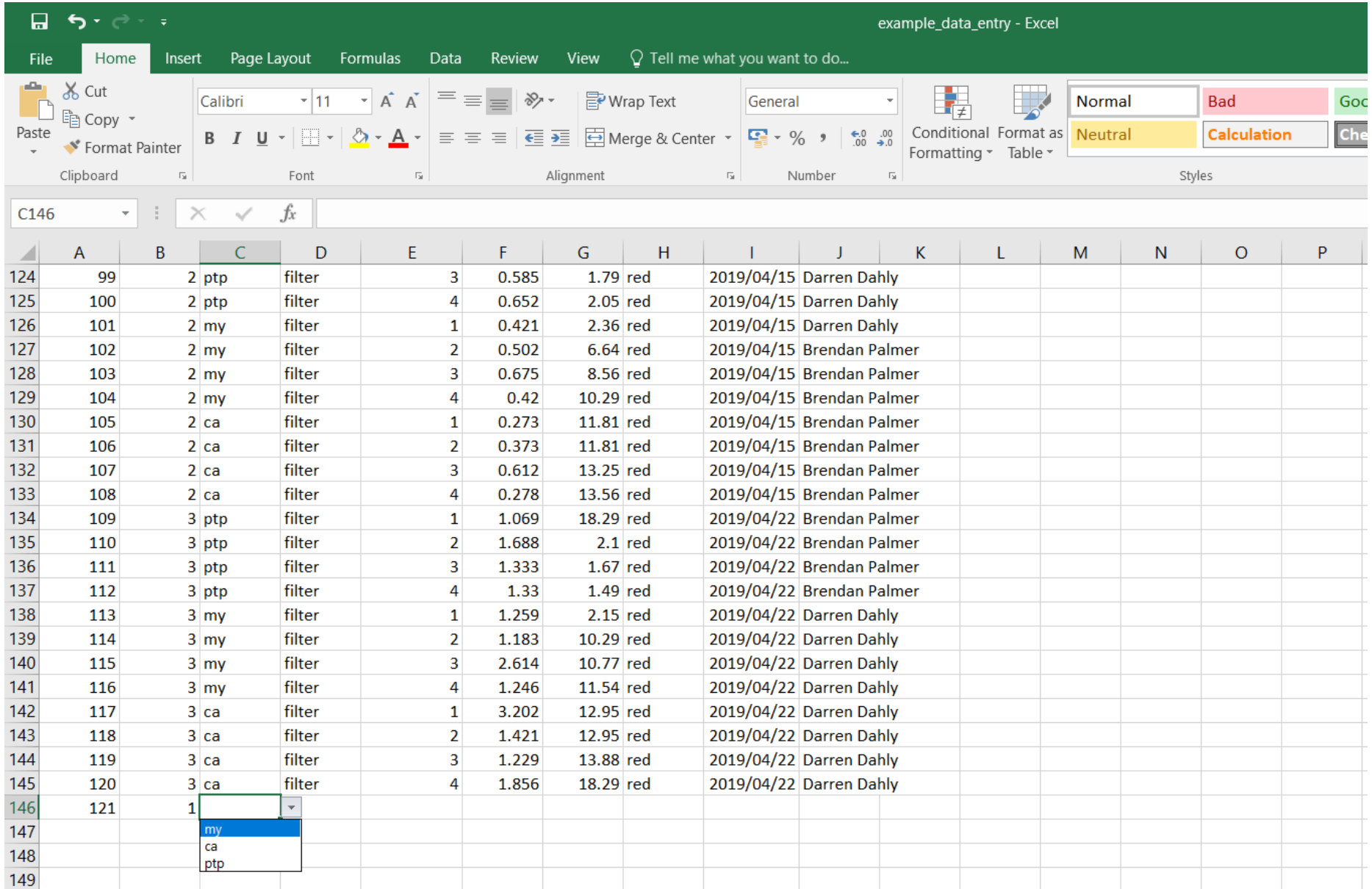
	A	B	C	D	E	
1	field_name	data_type	data_format	example	standard_units	description
2	id	numeric	integer	23	NA	Unique identifier applied to each observation
3	week_no	numeric	integer			
4	filter_name	character	NA	n		
5	treatment	character	NA	fi		
6	replicate_no	numeric	integer			
7	flavonoids	numeric	double			
8	biomass	numeric	double			
9	variety	character	NA	c		
10	date	date	YYYY/MM/DD			
11	investigator	character	Firstname Lastname	A		
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						
27						
28						
29						
30						

data | dictionary | values

	A	B	C	D	E	F	G	H	I	J	K
1	id	week_no	filter_name	treatment	replicate_no	flavonoids	biomass	variety	date	investigator	
2		0	my	filter	1			cos		Brendan Palmer	
3		1	ca	no_filter	2			sky		Darren Dahly	
4		2	ptp		3			red			
5		3			4						
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											
28											
29											
30											

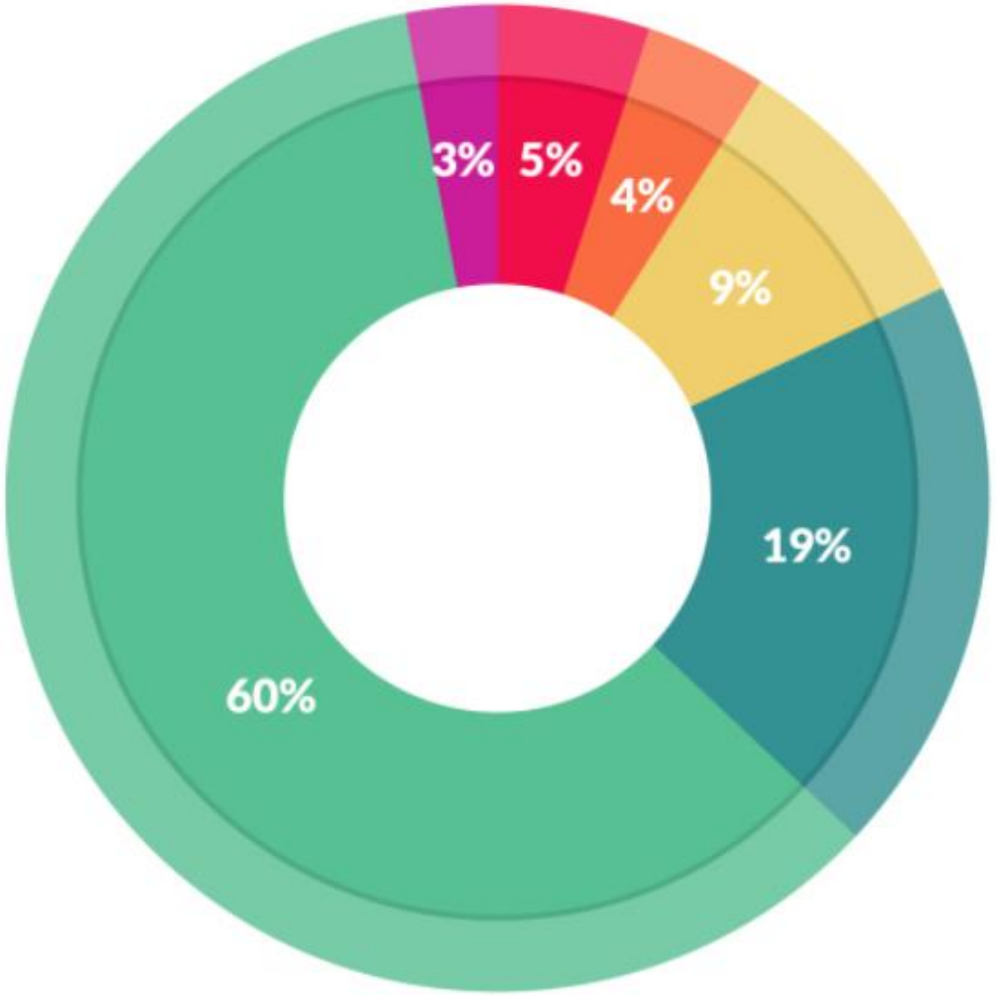
data | dictionary | values

Less stress, more success



	A	B	C	D	E	F	G	H	I	J
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								
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%*

CrowdFlower (2016) Data Science Report

Try it out yourself

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

Set status

Brendan Palmer

bapalmer

Edit profile

Overview Repositories 14 Projects 0 Stars 1 Followers 12 Following 10

Pinned Order updated. Customize your pins

RSS_Belfast_2019

Data FAIRification using R/RStudio workflows

R

R-A_Hitchhikers_Guide_to_Reproducible_Research

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

RCR

Section of the UCC Reproducible Conduct of Research digital badge dedicated to exposing researchers to reproducible research practices.

HTML

lunchtime_sessions

Short 1 hour introductions to R-related topics such as creating R projects, using GitHub through RStudio and more

HTML ★ 1

It doesn't get much easier than this!

holepunch_file.R*

Source on Save | Run | 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?...

Cold Spring Harbor Laboratory

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

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?].

Posted May 12, 2019.

Abstract | Full Text | Info/History | Metrics | Preview PDF

Download PDF

Data/Code

Email

Share

Citation Tools

Tweet Like 7

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.

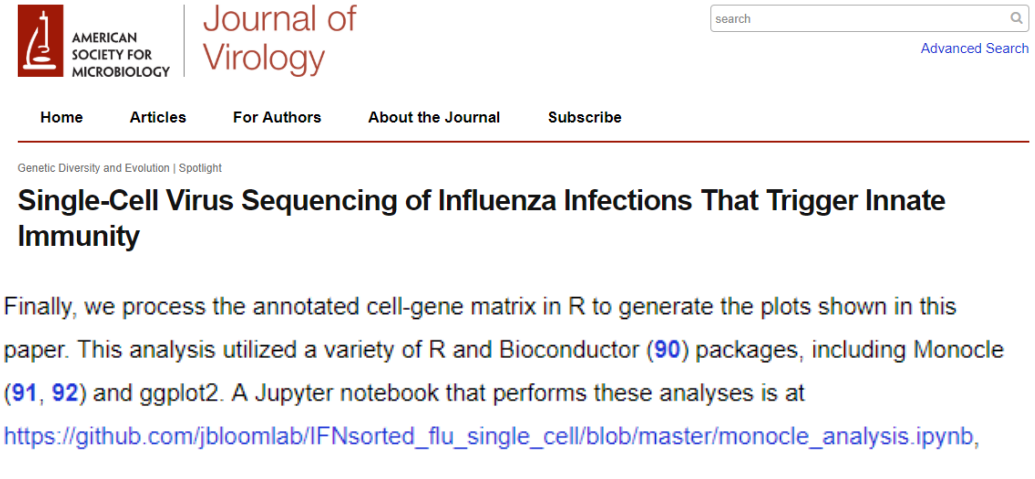
1999

* 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

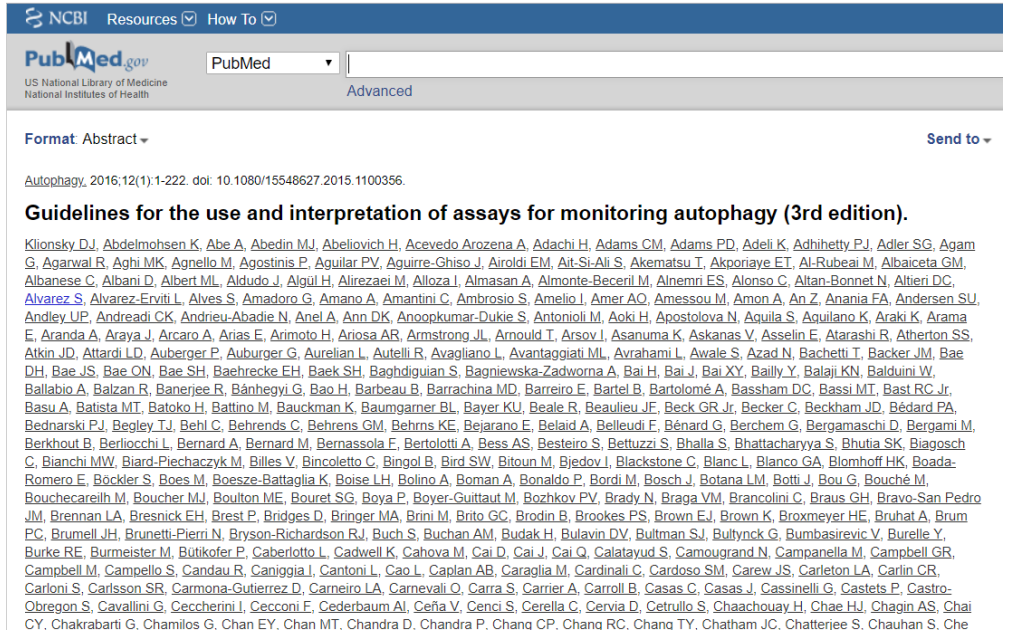


2019

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

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



jbloomlab / IFNsorted_flu_single_cell

Watch

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Security | Insights

Branch: master | IFNsorted_flu_single_cell / monocle_analysis.ipynb

jbloom saved plot of association between co-infection / IFN

1 contributor

12.8 MB

Table of Contents

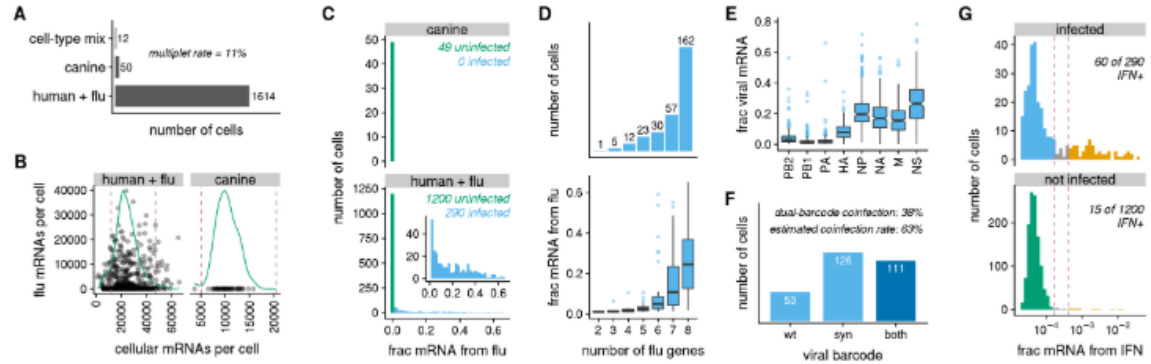
Figures for paper

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_flu_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_flu, labels="C", scale=0.95, label_size=18, vjust=1),
  plot_grid(p_nflu_genes, labels="D", scale=0.95, label_size=18, vjust=1),
  plot_grid(p_flu_rel_expr, p_coinfect,
            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)
```

Here today but maybe gone tomorrow?

Geoff Barton 🇪🇺
@gjbarton

We are applying to renew funding for core @Jalview development in @bartongrp. Please help us by writing a support letter to say how you find Jalview useful in your work. Send your letter as a PDF on headed paper by 15th Oct 2018 to: support_jalview@bartongroup.org. Thanks!

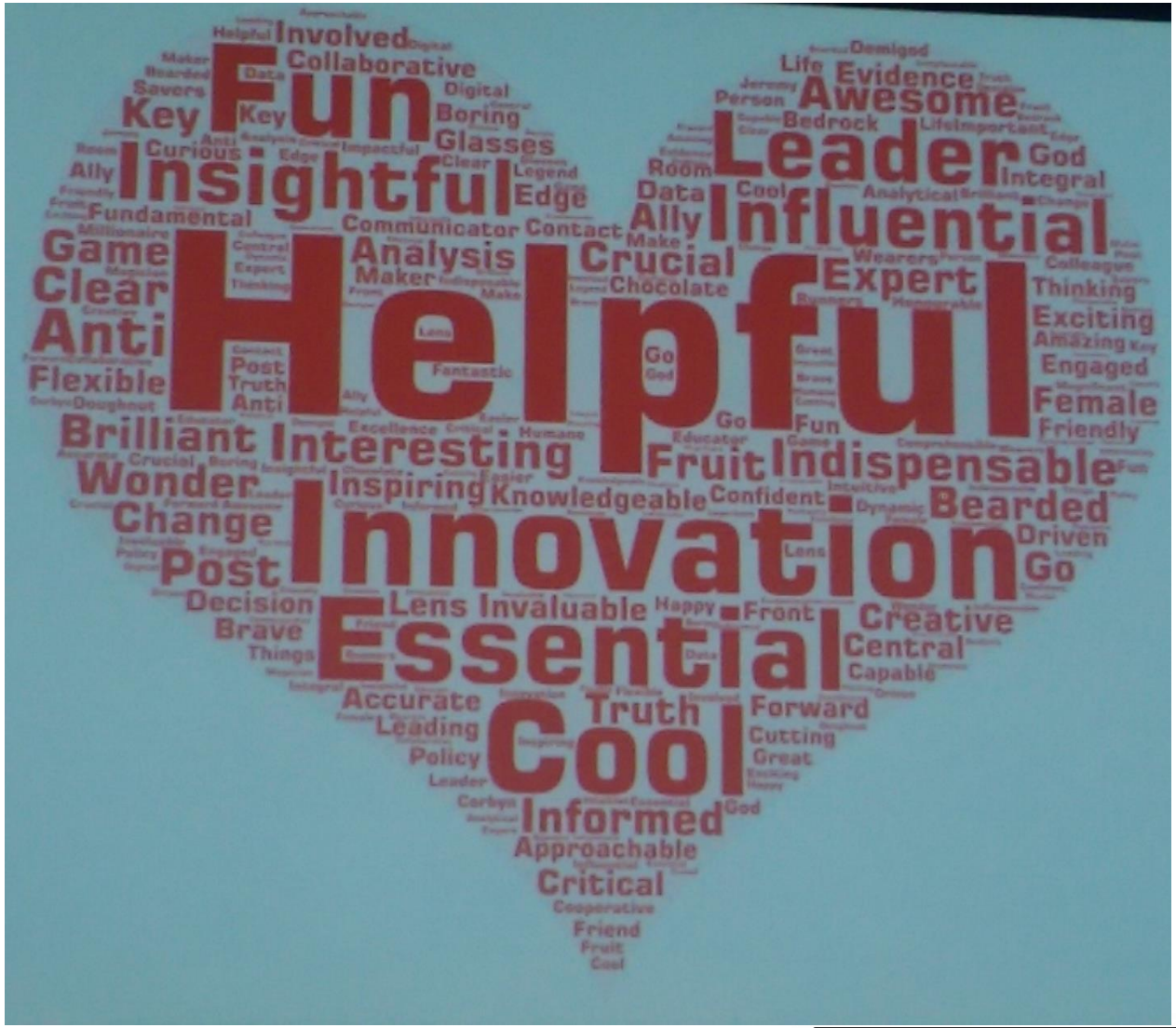
3:52 PM · Oct 3, 2018 · Twitter Web Client

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

Lots of inspiration to be found at #RSS2019Conf

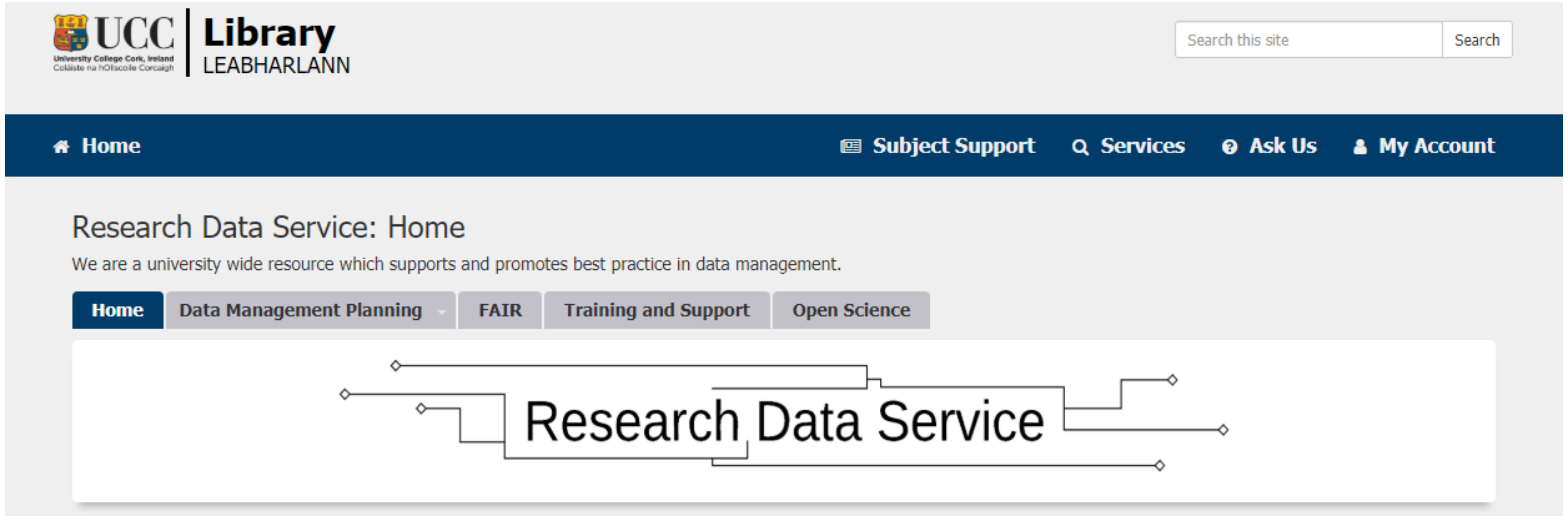


Siobhan Carey – Mondays Keynote

Acknowledgements/Local Supports

Dr Darren Dahly

Dr Brendan Palmer



Eoghan Ó Carragáin

Dr Aoife Coffee