Research Data Management

Dr. Sara El-Gebali

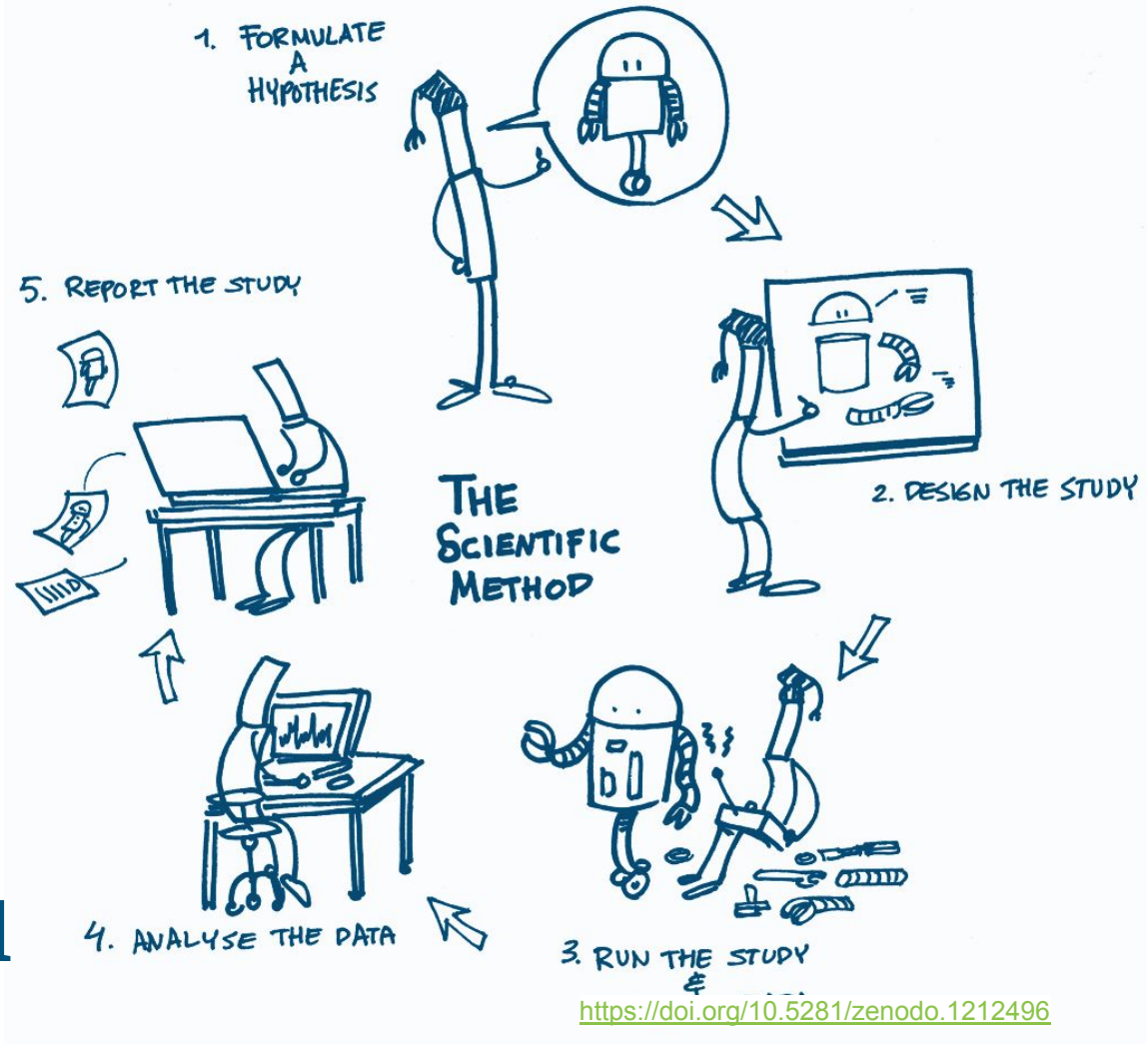
0000-0003-1378-5495
@yalahowy

26-April-2022

FAIR oints

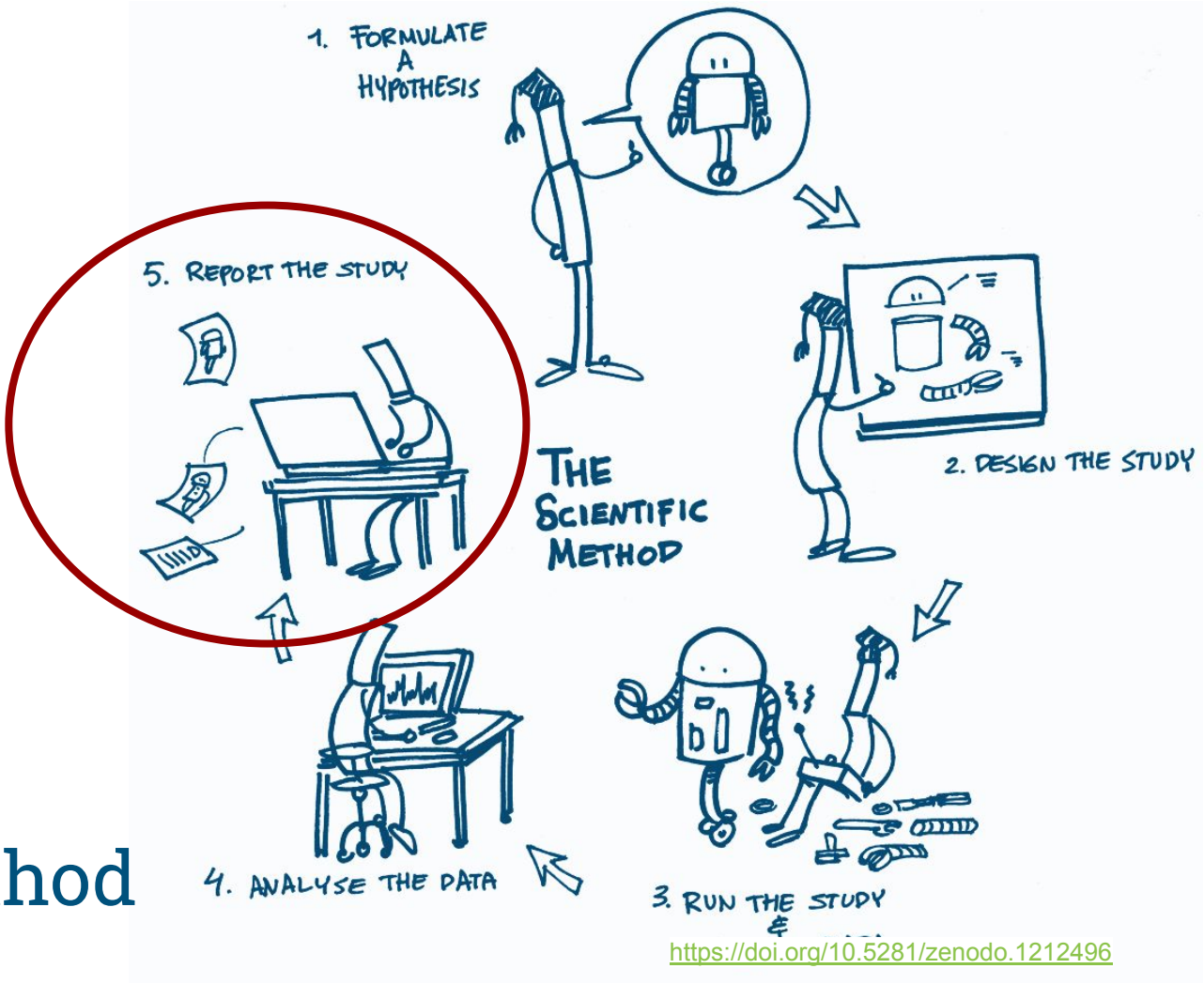
SciLifeLab

Scientific method



https://doi.org/10.5281/zenodo.1212496

Scientific method



https://doi.org/10.5281/zenodo.1212496

Open Data

Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike.

The Open Data Handbook- Open Knowledge Foundation

Why should make it Open?

Published: 25 April 1953

Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid

J. D. WATSON & F. H. C. CRICK

Nature **171**, 737–738 (1953) | Cite this article

153k Accesses | **8086** Citations | **1748** Altmetric | Metrics

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

This is a preview of subscription content

Access options

Access through your institution

Buy article

Get time limited or full article access on ReadCube.

$32.00

Buy

All prices are NET prices.

Subscribe to Journal

Get full journal access for 1 year

199,00 €

only 3,90 € per issue

Subscribe

Tax calculation will be finalised during checkout.

Why should make it Open?

Published: 25 April 1953

Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid

J. D. WATSON & F. H. C. CRICK

Nature **171**, 737–738 (1953) | Cite this article

153k Accesses | **8086** Citations | **1748** Altmetric | Metrics

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

This is a preview of subscription content

Access options

Access through your institution

Buy article

Get time limited or full article access on ReadCube.

$32.00

Buy

All prices are NET prices.

Subscribe to Journal

Get full journal access for 1 year

199,00 €

only 3,90 € per issue

Subscribe

Tax calculation will be finalised during checkout.

Open Science Beyond Open Access: For and with communities, A step towards the decolonization of knowledge

Chan, Leslie; Hall, Budd; Piron, Florence; Tandon, Rajesh; Williams, Wanósts'a7 Lorna

UNESCO is launching international consultations aimed at developing a Recommendation on Open Science for adop

https://zenodo.org/record/3946773#.YW_WiNIBzt0

Why should make it Open?

Open research data has accelerated investigations in pandemic research led to valuable discoveries

Sections

The Washington Post
Democracy Dies in Darkness

Get one year for £20

This article is more than 1 year old

Science

Scientists are unraveling the Chinese coronavirus with unprecedented speed and openness

Medical workers of the Union Hospital with the Tongji Medical College of the Huazhong University of Science and Technology in Wuhan participate in a ceremony to form an "assault team" to battle a coronavirus epidemic in China. (Cheng Min/Xinhua/AP)

By Carolyn Y. Johnson

January 24, 2020

Just 10 days after a pneumonia-like illness was first reported among people who visited a seafood market in Wuhan, China, scientists released the genetic sequence of the coronavirus that sickened them. That precious bit of data, freely available to any researcher who wanted to study it, unleashed a massive collaborative effort to understand the mysterious new pathogen that has been rapidly spreading in China and beyond.

The genome was posted on a Friday night on an open-access repository for genetic information. By Saturday morning, Andrew Mesecar, a professor of cancer structural biology at Purdue University, had redirected his laboratory to start analyzing the DNA sequence, which bore a striking resemblance to that of severe acute respiratory syndrome (SARS), the 2002 viral outbreak that sickened more than 8,000 people and killed nearly 800. Scientists at the National Institutes of Health's Rocky Mountain Laboratories in Montana asked a company to turn the information from a string of letters on a computer screen into actual DNA they could study in lab dishes.

Post REPORTS

Sound. Informed. Listen Now

"Just three weeks after the first viral sequence was published, more than 42 different genomes are available on Nextstrain, an open source viral genome database that continues to grow as scientists diagnose patients and publish the viral genomes in just a few days."

CBC | MENU

news | Top Stories | Local | The National | Opinion | World | Canada | Politics | Inc

Health · Second Opinion

'We're opening everything': Scientists share coronavirus data in unprecedented way to contain, treat disease

The current climate of sharing data is unusual for scientists, says researcher

Kelly Crowe · CBC News · Posted: Feb 01, 2020 4:00 AM ET | Last Updated: February 1

Medical staff in protective suits treat a patient with pneumonia caused by the coronavirus at the Zhongnan Hospital of Wuhan University in Wuhan, China, on Tuesday. (China Daily/Reuters)

Feb 01, Kelly Crowe · CBC News · Posted:, and 2020 4:00 AM ET | Last Updated: February 1. "'We're Opening Everything": Scientists Share Coronavirus Data in Unprecedented Way to Contain, Treat Disease | CBC News'. CBC. Accessed 4 February 2020. https://www.cbc.ca/news/health/coronavirus-2019-ncov-science-virus-genome-who-research-collaboration-1.5446948.

Hindawi

Is Open enough?

Enhancing access to research data during crises: lessons learned from the COVID-19 pandemic.

This is a background paper that has been prepared for an OECD Global Science Forum (GSF) workshop on 23 April, 2021, which is part of a broader project on *Mobilising science in response to COVID-19: lessons learned from COVID-19.*

a major obstacle to timely and effective access. Initiatives already underway in Europe and other regions to develop Open Science Clouds are not yet well enough developed to overcome this obstacle. In short, a lot of data that are extremely valuable for COVID-19 research and responding to the pandemic are not sufficiently, findable, accessible, interoperable and reusable (FAIR).

The unprecedented spread of the virus has prompted a rapid and massive research response and this has been greatly facilitated by well-established international data sharing initiatives - such as GISAID for SARS-CoV-2 genomic data. However, such initiatives remain restricted to certain research domains and in many fields there are no universally adopted systems or standards, for collecting, documenting and disseminating COVID-19 research data and associated code and software. Many data are not reusable by, or useful to, different communities if they have not been sufficiently documented and contextualised or appropriately licensed. This is not a new challenge for many areas of science but, in the context of COVID-19, it is a challenge that needs to be urgently addressed and for which, in many cases, solutions exist but have not been fully adopted.

The responsible, FAIR and timely sharing of data is an essential element of the Open Science approach that the world needs to effectively combat pandemics like COVID-19 and other complex crises. Unnecessarily limiting or

https://www.oecd.org/sti/inno/enhance-access-research-data-during-crises.htm

FAIR Data principles

nature > scientific data > comment > article

Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, [...] Barend Mons

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

An Addendum to this article was published on 19 March 2019

Abstract

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

FAIR Data principles

nature > scientific data > comment > article

Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, [...] Barend Mons

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

An Addendum to this article was published on 19 March 2019

Abstract

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

FAIR principles

FAIR is a set of principles to define the best practices for data & metadata to facilitate discovery, access and reuse by **humans** and **machines**.

FAIR is not rules and not a standard, it is an evolving process and a vision.

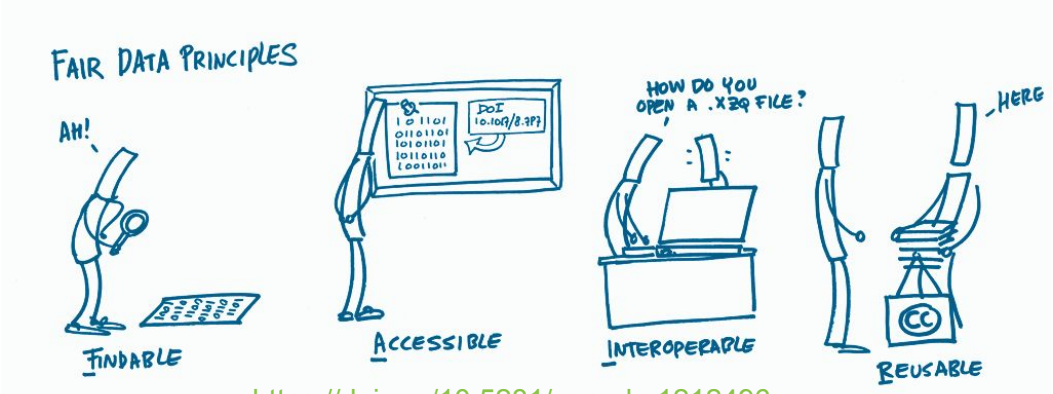
FAIR principles

FAIR is a set of principles to define the best practices for data & metadata to facilitate discovery, access and reuse by **humans** and **machines**.

FAIR is not rules and not a standard, it is an evolving process and a vision.

What does **FAIR** stand for?

Findable, **A**ccessible, **I**nteroperable and **R**eusable.



https://doi.org/10.5281/zenodo.1212496

FAIR Basics

The Four Basics of FAIR:

'Findable'	i.e. discoverable with metadata, identifiable and locatable by means of a standard identification mechanism
'Accessible'	i.e. always available and obtainable; even if the data is restricted, the metadata is open
'Interoperable'	i.e. both syntactically parseable and semantically understandable, allowing data exchange and reuse between researchers, institutions, organisations or countries; and
'Reusable'	i.e. sufficiently described and shared with the least restrictive licences, allowing the widest reuse possible and the least cumbersome integration with other data sources.

https://www.openaire.eu/how-to-make-your-data-fair

The term FAIR originates from the 2014 Lorentz Workshop resulting in 15 guiding principles published in 2016 to make research Findable, Accessible, Interoperable, and Reusable.

To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

F2. data are described with rich metadata.

F3. metadata specify the data identifier.

F4. (meta)data are registered or indexed in a searchable resource.

To be Accessible:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 the protocol is open, free, and universally implementable.

A1.2 the protocol allows for an authentication and authorization procedure, where necessary.

A2 metadata are accessible, even when the data are no longer available.

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles.

I3. (meta)data include qualified references to other (meta)data.

To be Re-usable:

R1. meta(data) have a plurality of accurate and relevant attributes.

R1.1. (meta)data are released with a clear and accessible data usage license.

R1.2. (meta)data are associated with their provenance.

R1.3. (meta)data meet domain-relevant community standards.

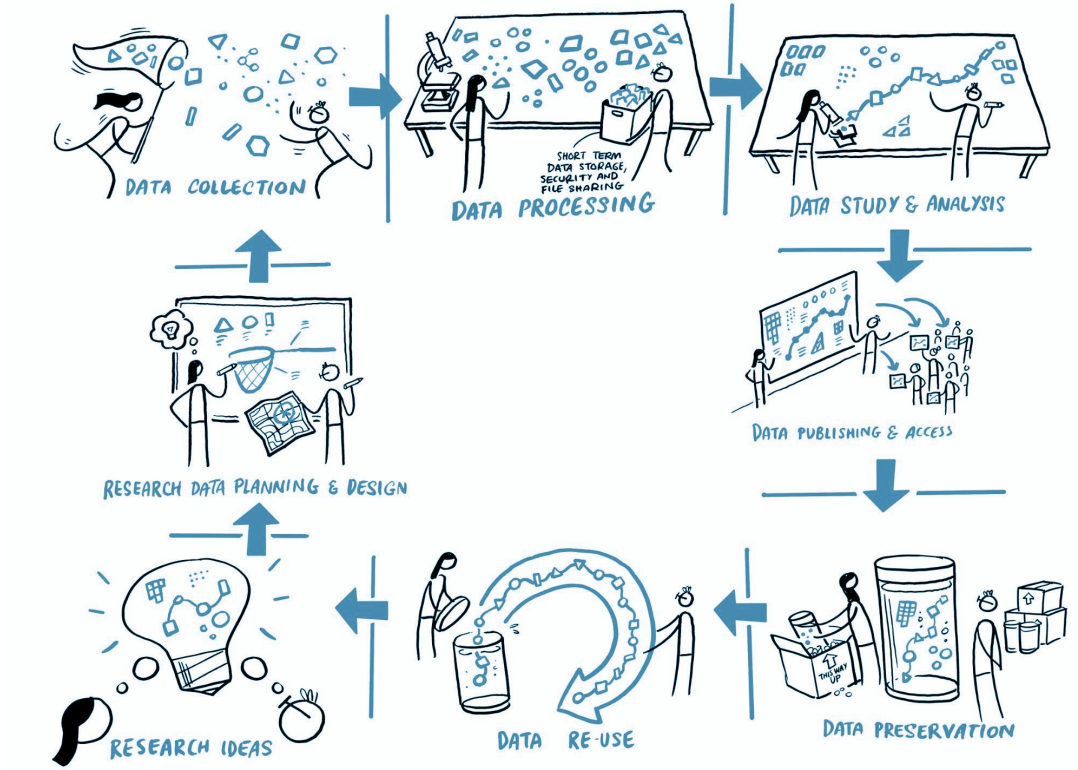
The latest developments on FAIR are available at GO-FAIR.

15 FAIR principles :
https://www.gofair.us/fair-principles

The path towards implementation

Research Data management

Everyday management of research data during the lifetime of a research project to preserve and share it beyond the project completion.



The Turing Way Community, & Scriberia. (2020, March 3). Illustrations from the Turing Way book dashes. Zenodo. http://doi.org/10.5281/zenodo.3695300

Research Data Lifecycle

Planning

- Data management planning (DMPs)
- Data description and metadata extraction
- Data documentation
- Choice of repositories
- Choices of file formats
- Data re-use
- Funders requirements
- File naming
- Ethics and Research conduct
- Funding for RDM activities

Managing

- Storage and backup & security
- Active Metadata collection
- Tools and software solutions
- Curation
- Versioning
- Provenance

Sharing

- Data access and Sharing rights
- Data privacy and GDPR compliance
- Data ownership, licensing
- Data Transfer

Preservation & Publication

- Citation
- PrePrint
- DOI
- Publishing requirements
- Long Term Storage
- Archival and Disposal policies

Findable

Your data should be findable, have appropriate description (i.e. metadata), have a persistent identifier.

Deposit the (meta)data in relevant repository with an assigned persistent identifier e.g. DOI or Handle

Domain-specific e.g. Registry of Research data Repositories (re3data)

General repositories e.g. Zenodo DataVerse

Personal identifier e.g. ORCID

Findable

Your data should be findable, have appropriate description (i.e. metadata), have a persistent identifier.

Deposit the (meta)data in relevant repository with an assigned persistent identifier e.g. DOI or Handle

Domain-specific e.g. Registry of Research data Repositories (re3data)

General repositories e.g. Zenodo DataVerse

Personal identifier e.g. ORCID

Annotate data with rich metadata using domain-agnostic or domain-specific controlled vocabularies

Domain-specific e.g. RDA Metadata Directory, Minimum Information for Biological and Biomedical Investigations (MIBBI)

General e.g. Dublin Core

More @FAIRsharing.org

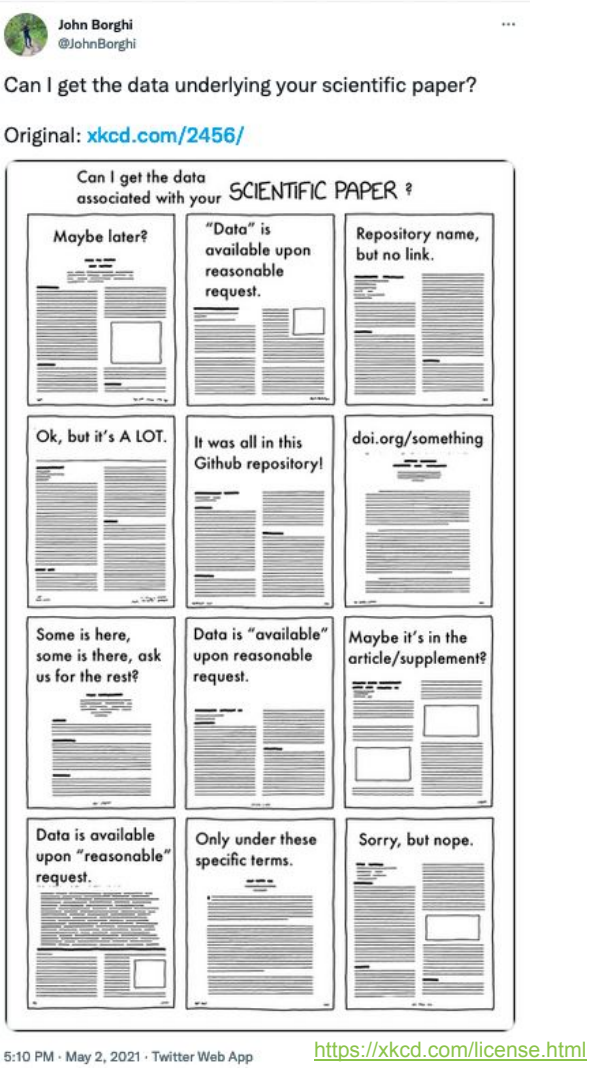
Accessible

Your data should be accessible for both humans and machines, i.e. retrievable and understandable

Accessible

Deposit the data under well defined conditions, i.e. data is accessible at **HTTP** or public **REST API**.

Specify what the users need to do to access this data, i.e. two factor authentication, request access from author, etc…



https://xkcd.com/license.html

Accessible

Your data should be accessible for both humans and machines, i.e. retrievable and understandable

Accessible ≠ Open

Deposit the data under well defined conditions, i.e. data is accessible at **HTTP** or public **REST API**.

Specify what the users need to do to access this data, i.e. two factor authentication, request access from author, etc…

Metadata should be made available and accessible;

'FAIR is not the equivalent of open, but open needs to be FAIR to be useful'

Remember

No license = No access!

‘As open as possible, as closed as necessary’

Even heavily protected and private data can be FAIR.

Interoperable

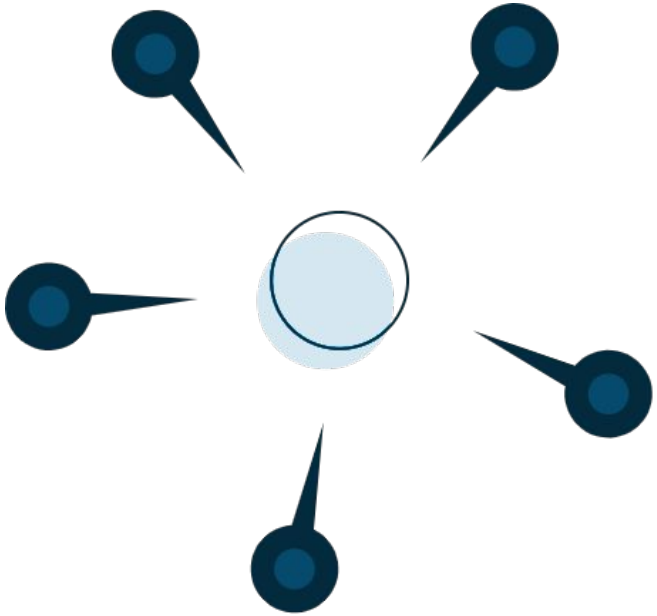
Machines and humans can interpret and use the data in different settings.

Rich metadata is key!

- **README** files

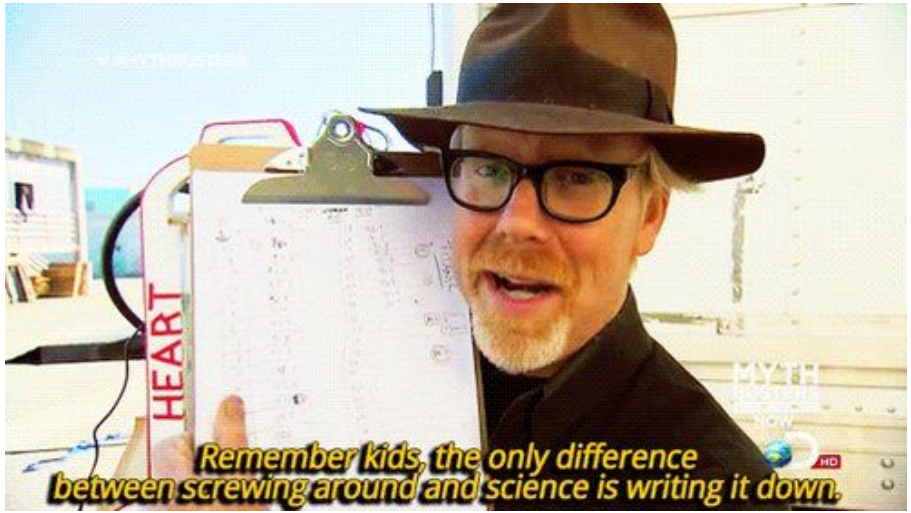
Describe your data properly, use controlled vocabulary, ontologies and standardise terminology.

Use preferred file formats, and open whenever possible.



Document everything!

- **Who** created and owns this data?
- **What** are the contents?
- **What** output and results?
- **When** was this data created and last updated?
- **Where** is it stored and published?
- **Which** methods were used?
- **Which** instruments were used?
- **How** was the data created, controlled and analysed?
- **How** can I use this data i.e. license?



https://www.tested.com/making/557288-origin-only-difference-between-screwing-around-and-science-writing-it-down/

Reuse

License

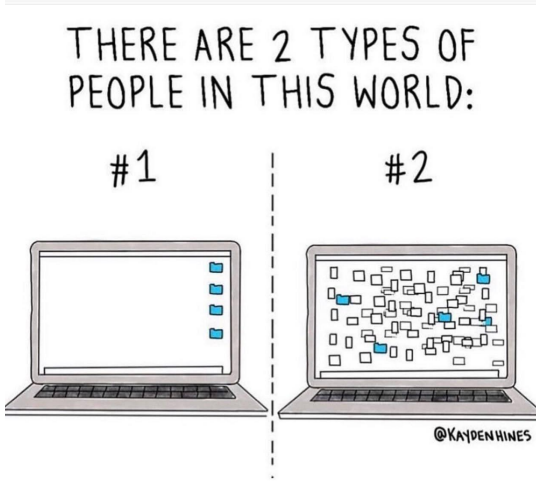
- License should be as open as possible but **NOT** necessary
- Add clear license, human and machine readable.
- Make sure it is inline with your institutional and funders requirements, intellectual property rights!!

Provenance

- Provenance information with controlled vocabularies
- Credit attribution
- How to cite

Build new habits

- Organize your data
- Name your files appropriately
- Choose file formats wisely
- Use versioning strategies
- Outline quality control strategies



FAIR Summary

1. Deposit your data where others can find it, keep in mind where your peers can find it, i.e. field specific repository and give it a stable unique identifier (PID).
2. Make your data & metadata accessible via standard means such as http/API.
3. Create metadata and explain in detail what this data is about, never assume people know!
4. Deposit metadata with PID and make it available with/out data i.e. in case data itself is heavily protected.
5. Include information on ownership, provenance and citation.
6. Outline what the reusers of your data are/not allowed to do, use clear license. Commonly used licenses like MIT or Creative Commons (keep in mind funders requirements).
7. Specify access conditions, if authentication or authorization is required.
8. Describe your data in a standardized fashion using agreed terminology and vocabulary.
9. Share the data in preferred & open file formats.

10. Start the process early on!!

FAIR ≠ Open

‘FAIR is not the equivalent of open, but open data needs to be FAIR to be useful’

Making your data freely and openly available does not translate to it being reusable!

To do so, we need clear, detailed contextual information and data description.

Data can be FAIR but not Open! FAIR data motto “as open as possible, as closed as necessary”

Ideally you want FAIR data shared openly!

Visit our website: https://www.fairpoints.org/

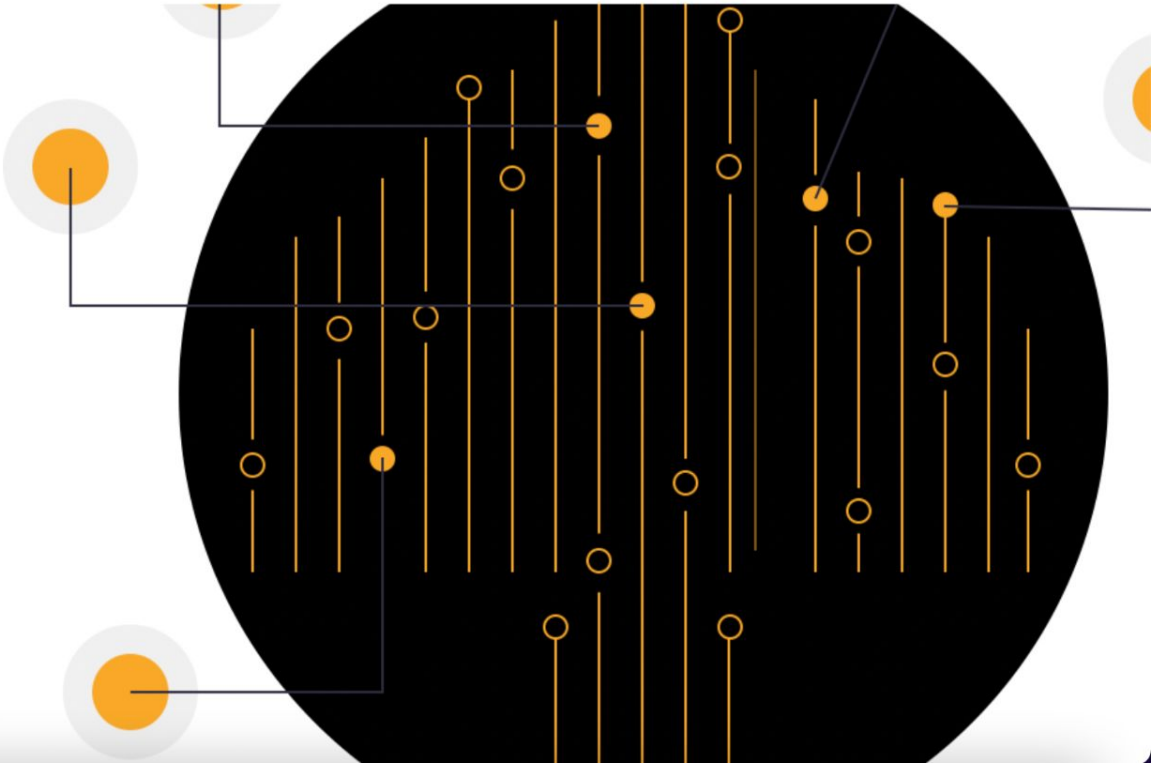
Home Events Get involved FAIRPoints resources Team Contact Participation Guidelines

FAIRPoints

The event series highlighting pragmatic measures developed by the community towards the implementation of the FAIR (Findable, Accessible, Interoperable, Reusable) data principles.

Email:
fairpoints@protonmail.com
Twitter FAIR_Points

CONTACT



Thank you!

Get in touch
Email: sara.elgebali@scilifelab.uu.se
Twitter: @yalahowy

FAIR software summary

1. Deposit in publicly accessible repositories https://software.ac.uk/choosing-repository-your-software-project
2. Use a version control system to easily track changes and versions; Github, Gitlab, Bitbucket,
3. Use of containers for software portability; Docker, Singularity
4. Describe with rich metadata including dependencies, with controlled vocabulary: Software Ontology, EDAM
5. Explain the intended use and conditions of functionality of the software
6. Add a license, Apache-2.0 and MIT are permissive licenses with few restrictions, allowing reuse. https://choosealicense.com/ // https://tldrlegal.com/
7. Register your code in a community https://github.com/NLeSC/awesome-research-software-registries
8. Store snapshots of your software with PIDs https://guides.github.com/activities/citable-code/
9. Enable proper citation for your software; CodeMeta and the Citation File Format were specifically designed to enable citation of software
10. **FAIR Software should operate on and deliver FAIR Data!**

Findable

F1. Register the software in relevant registry with an assigned DOI

General repositories such as Zenodo and Github

- **Language** specific; Python Package Index (PyPI) https://pypi.org/
- **Domain** specific; https://biocontainers.pro/

Findable

F1. Register the software in relevant registry with an assigned DOI

General repositories such as Zenodo and Github

- **Language** specific; Python Package Index (PyPI) https://pypi.org/
- **Domain** specific; https://biocontainers.pro/

F2. Annotate software using domain-agnostic or domain-specific controlled vocabularies

- **The Software Ontology**
- **EDAM-** Ontology of bioscientific data https://edamontology.org/page
- **OntoSoft**
- **More @FAIRsharing.org**

Findable

F3. Include software citation with metadata standards

- **The Citation File Format (CFF)**
- **A CodeMeta instance file**
- **Biotools Schema**
- **Bioschemas Tool profile**

Findable

F3. Include software citation with metadata standards

- **The Citation File Format (CFF)**
- **A CodeMeta instance file**
- **Biotools Schema**
- **Bioschemas Tool profile**

CiteAs[1]
alpha

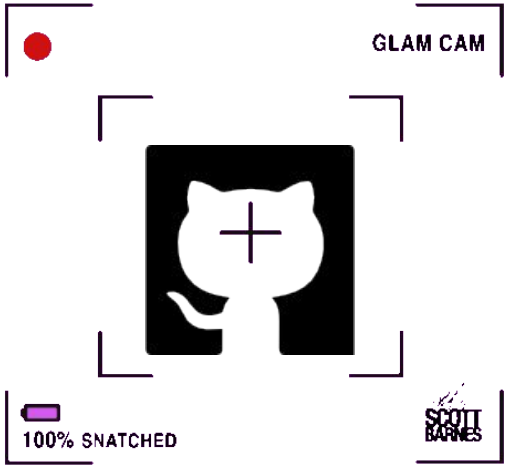
All research products deserve credit.

Get the correct citation for diverse research products, from software and datasets to preprints and articles.

Paste a URL, DOI, arXiv ID, or any search term (e.g. software name/abbreviation)

Examples: http://yt-project.org https://cran.r-project.org/web/packages/stringr More examples

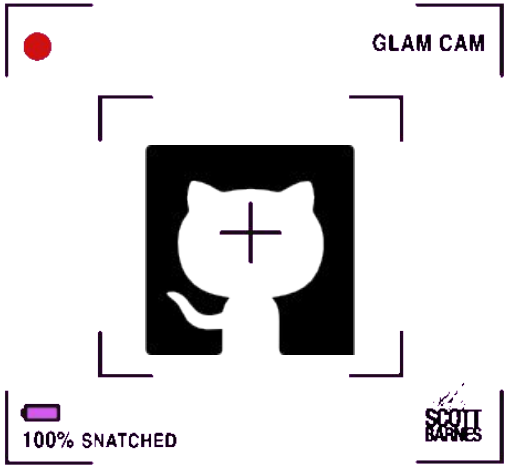
Accessible



Take a snapshot from Github

Software stored on Github is accessible for use, reuse and allows for engagement with the community and **versioning**.

Accessible

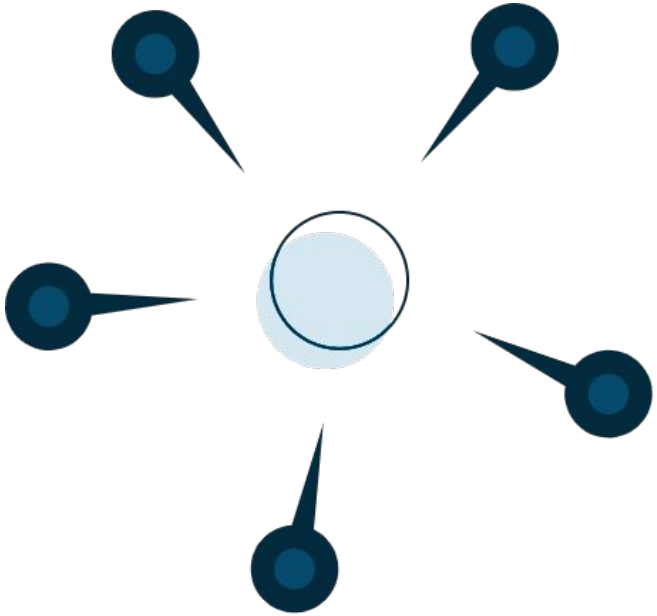


Take a snapshot from Github
Software stored on Github is accessible for use, reuse and allows for engagement with the community and **versioning**.

Deposit in Zenodo
Zenodo offers archival (~20 years), **PID** and opportunity for reproducibility.

Interoperable

- Rich metadata is key!
- The use of Common Workflow Language (CWL), or Workflow Description Language (WDL) enables the interoperability between different pieces of software and workflow platforms
- Containers (e.g. use Docker, singularity) allows for accessibility across different operating systems and environments i.e. software portability.



Reuse

License

- License should be as open as possible
- Add clear license, human and machine readable e.g. Software Package Data Exchange standard
- License of software components should be compatible

REUSE SOFTWARE

Get Started | FAQ | Developers | Specification | Resources | Supporters | API

REUSE SOFTWARE

We make licensing easy for humans and machines alike. We solve a fundamental issue that Free Software licensing has at the very source: what license is a file licensed under, and who owns the copyright? **Adopting our recommendations is as easy as one-two-three**!

REUSE SOFTWARE

1. Choose and provide licenses
2. Add copyright and licensing information to each file
3. Confirm REUSE compliance

Reuse

License

- License should be as open as possible
- Add clear license, human and machine readable e.g. Software Package Data Exchange standard
- License of software components should be compatible

Provenance

- Provenance information with controlled vocabularies e.g. PROV-O
- Credit attribution
- How to cite and contribute

Resources

Resources- Research data management:

- Research Data Management 1 day workshop
- Making the Case for Research Data Management
- What is Research Data?
- New England Collaborative Data Management Curriculum
- Science Europe- RDM guide
- Roberts Lab Handbook- Data management in life sciences
- Research data management (RDM) open training materials

Resources- What is Data & FAIR data:

- Research Libraries
- Zenodo-FAIR principles
- “A love letter to your future self”: What scientists need to know about FAIR data
- Invest 5% of research funds in ensuring data are reusable
- FAIRaware- FAIR assessment
- H2020 Programme Guidelines on FAIR Data Management in Horizon 2020
- FAIRsFAIR Europe
- How to FAIR
- Go FAIR
- FAIR sharing

Resources- Open Data & reuse, reproducibility:

- Mozilla open science- challenges to open data and how to respond
- Ten arguments against Open Science that you can win
- ‘I ain’t afraid of no myth’ – busting the myths on data sharing
- OpenAIRE Research Data Management Hand Book
- Open Data Hand Book
- Open Science Hand Book
- Papers Without Code - submission
- Open Data and FAIR Data: differences and similarities
- Open Science Mooc
- Rein in the four horsemen of irreproducibility
- Making experimental data tables in the life sciences more FAIR: a pragmatic approach
- Open Scientist Handbook
- The Turing Way

Resources- Data Organization:

- Towards a Standardized Research Folder Structure
- Swedish National Data service
- DataOne research data management modules
- Imperial College Research data management guides
- King’s college- Managing your data
- UK Data Services

Resources- Software

- Reproducible analysis and Research Transparency
- Data Science for the Biomedical Sciences
- From FAIR research data toward FAIR and open research software
- Towards FAIR principles for research software
- Software vs. data in the context of citation
- Assessment report on ‘FAIRness of software’
- Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv
- FAIR Software
- Python package to analyze a GitHub or GitLab repository’s compliance with the fair-software.eu recommendations.
- Checklist for a Software Management Plan
- Top 10 FAIR Data & Software Things
- Research Software Alliance
- CodeMeta
- Open source for open science- CERN
- Software Reproducibility - The Nuts and Bolts
- Is software reproducibility possible and practical?
- Make a README
- README awesome list