

Open Science & FAIR

Introduction to Data Management Practices course

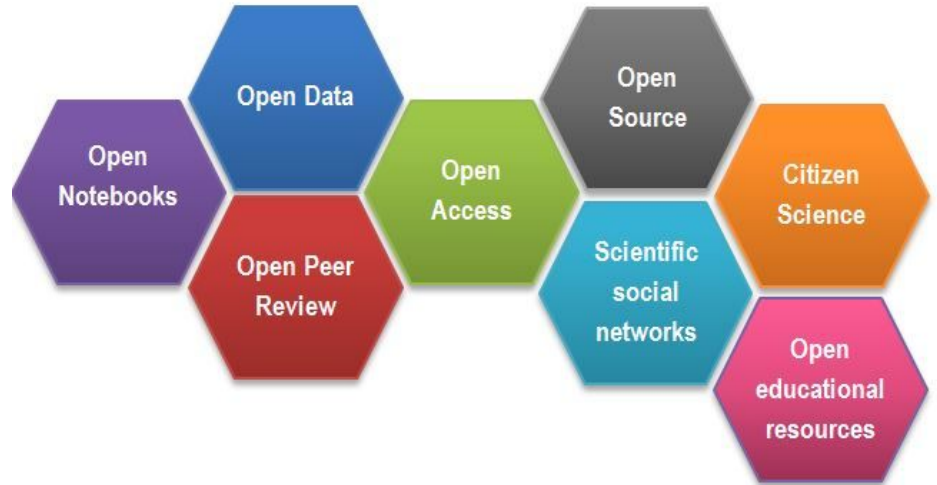
NBIS DM Team

data-management@scilifelab.se

Open Science

Make scientific research and its dissemination accessible to all levels of society.

- Open methodology
- **Open source**
- **Open data**
- Open access
- Open peer review
- Open educational resources



"Open Science facets as a beehive" by Gema Bueno de la Fuente licenced under [CC-BY](#)

What do you think are reasons for Open Data?

Open Data

- Democracy and transparency
 - Publicly funded research data should be accessible to all
 - Published results and conclusions should be possible to check by others
- Research
 - Enables others to combine data, address new questions, and develop new analytical methods
 - Reduce duplication and waste
- Innovation and utilization outside research
 - Public authorities, companies, and private persons outside research can make use of the data
- Citation
 - Citation of data will be a merit for the researcher that produced it



Ethical?

Doing “sloppy” science & not being open and transparent

Waste of resources

Contributing to the current research credibility crisis

Contributing to the current reproducibility crisis

Harming the profession

Harming public trust in research

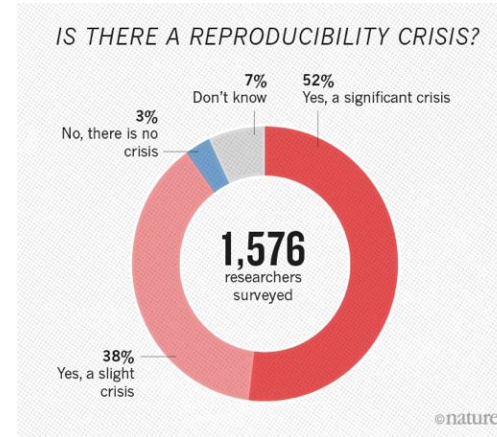
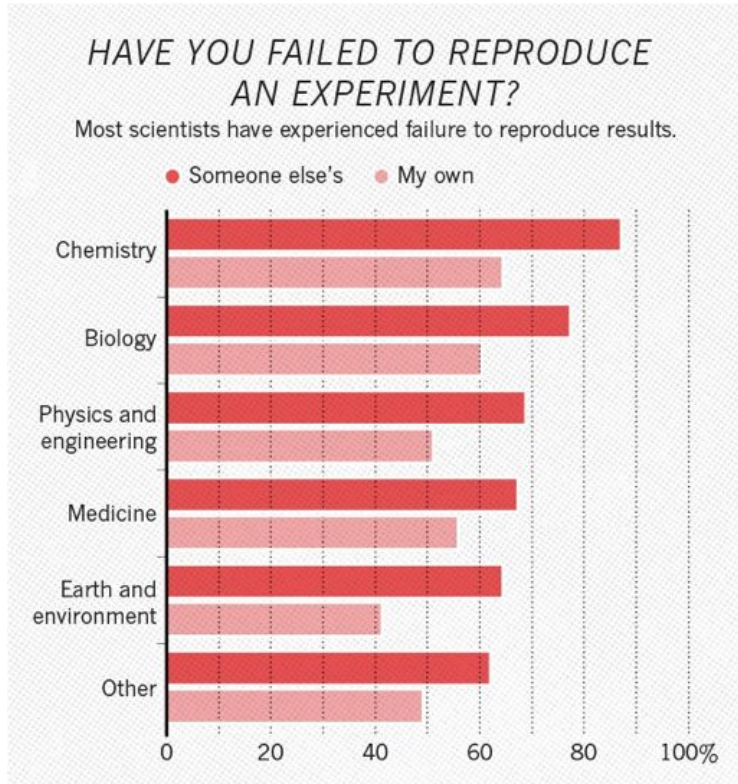


Crisis?

Do you think we have a **credibility** and/or
reproducibility crisis?

If so, what are some of its causes?

A reproducibility crisis



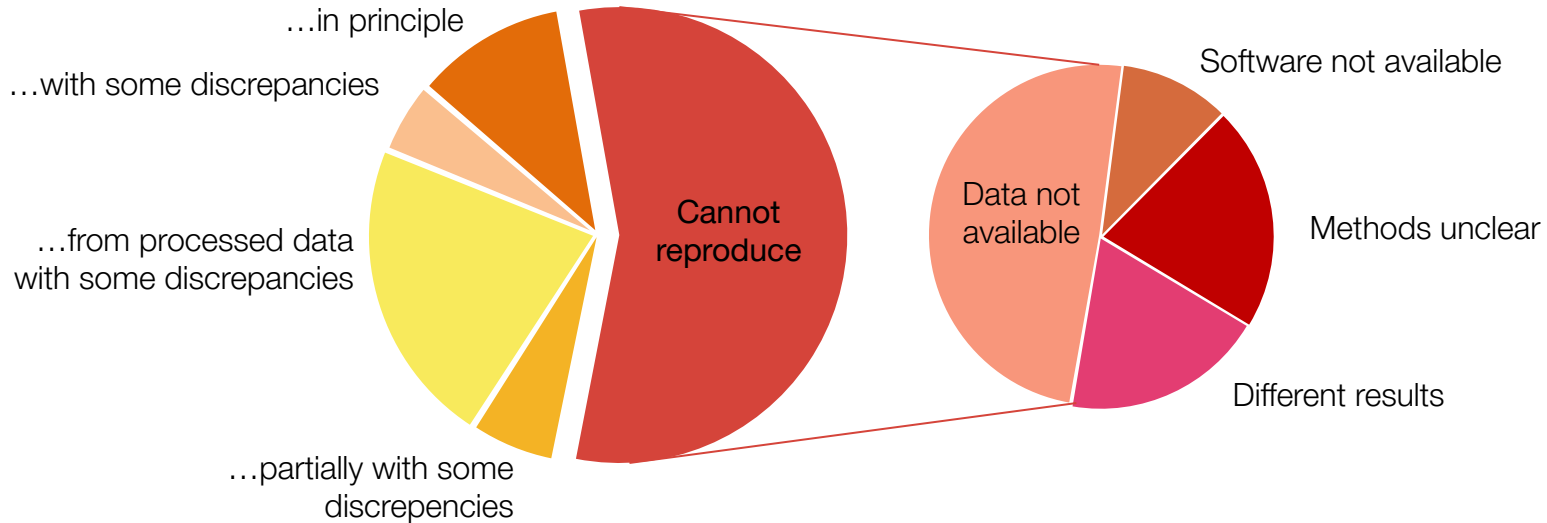
[1] "1,500 scientists lift the lid on reproducibility". Nature. 533: 452–454

[2] Begley, C. G.; Ellis, L. M. (2012). "Drug development: Raise standards for preclinical cancer research". Nature. 483 (7391): 531–533.

A reproducibility crisis

Reproduction of data analyses in 18 articles on microarray-based gene expression profiling published in *Nature Genetics* in 2005–2006:

Can reproduce...



Summary of the efforts to replicate the published analyses.

Adopted from: Ioannidis et al. Repeatability of published microarray gene expression analyses.

Nature Genetics 41 (2009) doi:10.1038/ng.295

Data Management Snafu



<https://www.youtube.com/watch?v=N2zK3sAtr-4>

FAIR

- To be useful for others data should be
 - **FAIR** - Findable, Accessible, Interoperable, and Reusable
... for both Machines and Humans

Wilkinson, Mark et al. “*The FAIR Guiding Principles for scientific data management and stewardship*”.
Scientific Data 3, Article number: 160018 (2016) <http://dx.doi.org/10.1038/sdata.2016.18>

The image shows the cover of a scientific article. At the top, the URL 'www.nature.com/scientificdata' is visible. The title 'SCIENTIFIC DATA' is prominently displayed in a large, blue, sans-serif font, with a decorative graphic of binary code (0s and 1s) to its right. Below the title, the word 'OPEN' is written in orange, followed by the article title 'Comment: The FAIR Guiding Principles for scientific data management and stewardship' in blue. The authors 'Mark D. Wilkinson et al.*' are listed below. A short abstract follows, discussing the need for improved data infrastructure and the role of the FAIR principles. At the bottom, there is a section titled 'Supporting discovery through good data management' which explains that good data management is a means to an end, leading to knowledge discovery and innovation.

www.nature.com/scientificdata

SCIENTIFIC DATA

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

Mark D. Wilkinson et al.*

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. Infrastructure, the enabling digital environment

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

Box 2 | The FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- 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 Reusable:

- R1. (meta)data are richly described with 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 detailed provenance
 - R1.3. (meta)data meet domain-relevant community standards



Findable

- Data have a **globally unique persistent identifier**
 - *e.g. a DOI, database accession number, etc*
- Data are described by **metadata**
 - *Information that explains the data*
- Data and metadata are findable in a **search resource**
 - *There must be ways of searching for the data*

Accessible

- Data is retrievable through a **standardised communication protocol** (open, free, allowing authentication & authorisation where necessary)
 - *e.g. http, sftp, etc*
- Metadata are accessible, **even if data is no longer available**
 - *Information about the data can be found even if data is no longer available*

Interoperable

- Metadata use a formal, accessible, shared **language for knowledge representation**
 - *Metadata is available in a form that even a computer can make use of*
- Metadata use **vocabularies** that follow the FAIR principles
 - *Standardised ways of capturing information about the data (that are in themselves FAIR)*
- Metadata include qualified **references** to other metadata
 - *If the data relies on other data, there must be links to those*

Reusable

- Data have a clear **data usage license**
 - *It is obvious under what conditions the data can be reused*
- Metadata are associated with **detailed provenance**
 - *The metadata is detailed enough to understand for what research questions it is relevant to reuse*
- Metadata meet domain-relevant community **standards**
 - *Metadata is described according to existing standards in the research field*

FAIR

- Both humans and machines are intended users of data
- The principles are not necessarily about *open* data
 - “As open as possible, as closed as necessary”
- FAIRness is not something absolute
 - Different levels of FAIR maturity
- FAIR does not enforce any particular technical standards

Simple FAIR data example



ArrayExpress

Search: "organism:mus_musculus" AND exptype:"RNA-seq of cod" 

Examples: [E-MEXP-31](#), [cancer](#), [p53](#), [Geuvadis](#)  [advanced search](#)










Home | Browse | Submit | Help | About ArrayExpress Contact Us |  Login

Filter search results


Show more data from EMBL-EBI

Search results for "organism:mus_musculus" AND exptype:"RNA-seq of coding RNA" AND expdesign:"time series" AND "organism part:liver"

6 experiments

Accession	Title	Type	Organism	Assays	Released	Processed	Raw	Atlas
E-MTAB-10239	scRNA-seq of murine mucosal associated invariant T (MAIT) cells after Francisella tularensis infection	RNA-seq of coding RNA from single cells	Mus musculus	3	19/05/2021	-		-
E-MTAB-7054	Transcriptional profiling of hepatic stellate cells (HSCs) isolated from Western diet/high fructose-fed C57BL6/J mice, carbon tetrachloride (CCl4)-treated C57BL6/J mice, and of murine HSCs differentiated in vitro	RNA-seq of coding RNA	Mus musculus	53	07/04/2019			-
E-MTAB-6435	Transcriptome profiling of liver samples of C/EBPβ ΔuORF mice	RNA-seq of coding RNA	Mus musculus	24	16/01/2019	-		
E-MTAB-7020	RNA-seq study on time of day specific Glucocorticoid action in mouse liver and lung tissues	RNA-seq of coding RNA	Mus musculus	32	13/11/2018	-		-
E-MTAB-7017	RNASeq data analysis of wild type and revert alpha knockout cells from mouse liver, at different time points, with or without DEX treatment	RNA-seq of coding RNA	Mus musculus	40	13/11/2018	-		
E-MTAB-2351	RNA-seq of Sod1 deficient and wild type mice after lymphocytic choriomeningitis virus (LCMV) infection	RNA-seq of coding RNA	Mus musculus	18	17/11/2015	-		-

 [Export table in Tab-delimited format](#)

 [Export matching metadata in XML format](#)

 [Subscribe to RSS feed matching this search](#)

Picture source: [ArrayExpress @ EMBL-EBI](#)



When to be FAIR?

FAIR at source?



Retroactively?

Good Data Management Practices

- **Data Management Plans**, to do your thinking ahead of time
- **Using standard metadata descriptions**, to clearly define your data
- **Organising your analysis**, so you and others can understand what you have done
- **Use versioning control** to keep track of changes you do
- **Clean up metadata and data** to be consistent with the standards you have chosen
- **Submit your data to international public repositories**, so others can find and reuse your data
- **Use scripted analysis of your data**, that can be understood by others

What data management practices do you apply in your research projects today?

	Ad Hoc	One-Time	Active and Informative	Optimized for Re-Use
Planning your project	When it comes to my data, I have a "way of doing things" but no standard or documented plans.	I create some formal plans about how I will manage my data, but I generally don't refer back to them.	I develop detailed plans about how I will manage my data that I actively revisit and revise over the course of a project.	I design my plans for managing data to streamline future use by myself or others.
Organizing your data	I don't follow a consistent approach for keeping my data organized, so it often takes time to find things.	I have an approach for organizing my data, but I only put it into action after my project is complete.	I have an approach for organizing my data that I implement prospectively, but it not necessarily standardized.	I organize my data to the so that others can navigate, understand, and use it without me being present.
Saving and backing up your data	I decide what data is important while I am working on it and typically save it in a single location.	I know what data needs to be saved and I back it up after I'm done working on it to reduce the risk of loss.	I have a system for regularly saving important data while I am working on it. I have multiple backups.	I save my data in a manner and location designed maximize opportunities for re-use by myself and others.
Getting your data ready for analysis	I don't have a standardized or well documented process for preparing my data for analysis.	I have thought about how I will need to prepare my data, but I handle each case in a different manner.	My process for preparing data is standardized and well documented.	I prepare my data in such a way as to facilitate use by both myself and others in the future.
Analyzing your data and handling the outputs	I often have to redo my analyses or examine their products to determine what procedures or parameters were applied.	After I finish my analysis, I document the specific parameters, procedures, and protocols applied.	I regularly report the specifics of both my analysis workflow and decision making process while I am analyzing my data.	I have ensured that the specifics of my analysis workflow and decision making process can be put into action by others.
Sharing and publishing your data	I share the results of my research, but generally I do not share the underlying data.	I share my my data only when I'm required to do so or in response to direct requests from other researchers.	I regularly share the data that underlies my results and conclusions in a form that enables use by others.	Because of my excellent data management practices, I am able to efficiently share my data whenever I need to with whomever I need to.

Borghi, J. et al (2018). *Support your Data*.
<https://doi.org/10.3897/rio.4.e26439>



Voting

	Ad Hoc	One-Time	Active and Informative	Optimized for Re-Use
Planning your project	When it comes to my data, I have a "way of doing things" but no standard or documented plans.	I create some formal plans about how I will manage my data, but I generally don't refer back to them.	I develop detailed plans about how I will manage my data that I actively revisit and revise over the course of a project.	I design my plans for managing data to streamline future use by myself or others.
Organizing your data	I don't follow a consistent approach for keeping my data organized, so it often takes time to find things.	I have an approach for organizing my data, but I only put it into action after my project is complete.	I have an approach for organizing my data that I implement prospectively, but it not necessarily standardized.	I organize my data to the so that others can navigate, understand, and use it without me being present.
Saving and backing up your data	I decide what data is important while I am working on it and typically save it in a single location.	I know what data needs to be saved and I back it up after I'm done working on it to reduce the risk of loss.	I have a system for regularly saving important data while I am working on it. I have multiple backups.	I save my data in a manner and location designed maximize opportunities for re-use by myself and others.
Getting your data ready for analysis	I don't have a standardized or well documented process for preparing my data for analysis.	I have thought about how I will need to prepare my data, but I handle each case in a different manner.	My process for preparing data is standardized and well documented.	I prepare my data in such a way as to facilitate use by both myself and others in the future.
Analyzing your data and handling the outputs	I often have to redo my analyses or examine their products to determine what procedures or parameters were applied.	After I finish my analysis, I document the specific parameters, procedures, and protocols applied.	I regularly report the specifics of both my analysis workflow and decision making process while I am analyzing my data.	I have ensured that the specifics of my analysis workflow and decision making process can be put into action by others.
Sharing and publishing your data	I share the results of my research, but generally I do not share the underlying data.	I share my my data only when I'm required to do so or in response to direct requests from other researchers.	I regularly share the data that underlies my results and conclusions in a form that enables use by others.	Because of my excellent data management practices, I am able to efficiently share my data whenever I need to with whomever I need to.

https://bit.ly/support_your_data_rubric

The Political landscape



'We support appropriate efforts to promote open science and facilitate appropriate access to publicly funded research results on findable, accessible, interoperable and reusable (FAIR)'

The Political Landscape

- Policymakers are **pushing for research data to be made available** as openly as possible
- Big investments are being made in **infrastructure and skills for data sharing and reuse**



**EUROPEAN OPEN
SCIENCE CLOUD**



SND

Swedish National Data Service

Policy landscape



- [UNESCO's Recommendation on Open Science](#)
- EU
 - Open Science Policy
 - Directives
 - European Research Council
 - Horizon Europe
 - EOSC
- Swedish research bills 2016 & 2020
 - Transition to open research data implemented by 2026
 - Government assignments to KB & VR
- [SUHF national roadmap for open science](#)
- [University policies](#)
- [Lund Declaration on Maximising the Benefits of Research Data](#)
- [National guidelines for open science](#) (KB)

Open Science **FAIR**

*“FAIR [...] open data sharing
should become the default [...]”*

*“As open as possible,
as closed as necessary”*



Motivators



Good Data Management Practices

- **Data Management Plans**, to do your thinking ahead of time
- **Using standard metadata descriptions**, to clearly define your data
- **Organising your analysis**, so you and others can understand what you have done
- **Use versioning control** to keep track of changes you do
- **Clean up metadata and data** to be consistent with the standards you have chosen
- **Submit your data to international public repositories**, so others can find and reuse your data
- **Use scripted analysis of your data**, that can be understood by others