

# Open Science & FAIR

*Introduction to Data Management Practices course*

NBIS DM Team

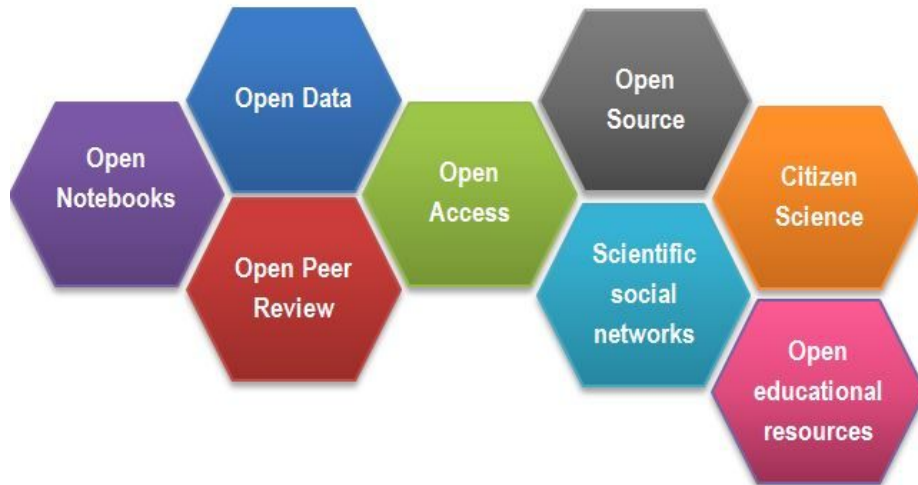
data-management@scilifelab.se

<https://doi.org/10.17044/scilifelab.c.6820587>



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



**What do you think are reasons for 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



---

*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*

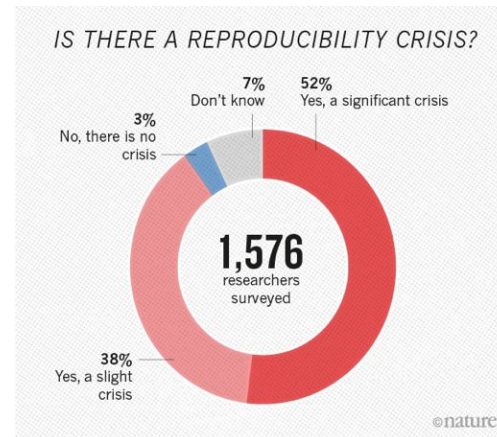
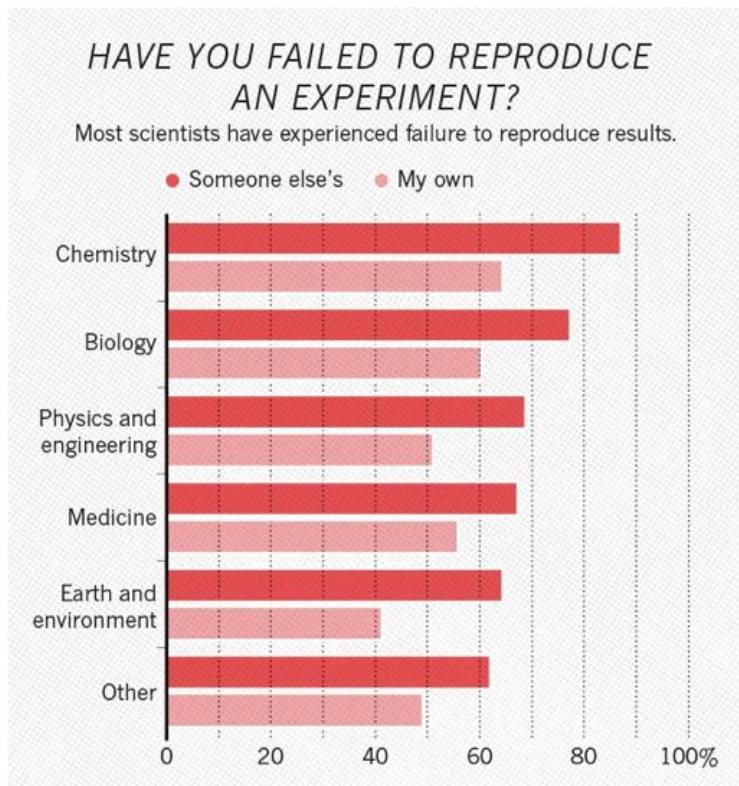
*Harming public trust in research*

My take of material by Rochelle Tractenberg “[Unexpected Ethical Challenges in Bioinformatics and Genomics.](#)”

---

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

If so, what are some of its causes?

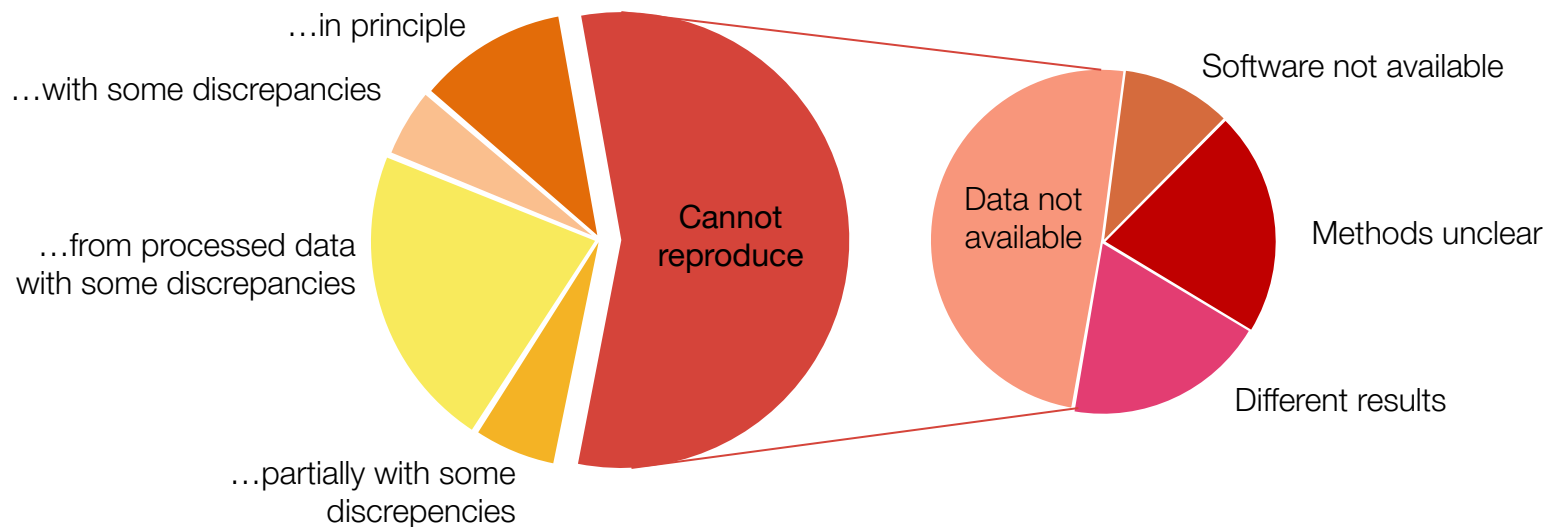


[1] "1,500 scientists lift the lid on reproducibility". Nature. 533: 452–454

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

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





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

- 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>



## 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

## *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

- 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*

- 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*

- 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*

- 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*

- 
- 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












## Find:

- **RNA-sequencing** experiments,
- from **liver** tissue,
- from the **common house mouse**,
- with a **time series** design



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
<a href="#">E-MTAB-10239</a>	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	-		-
<a href="#">E-MTAB-7054</a>	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			-
<a href="#">E-MTAB-6435</a>	Transcriptome profiling of liver samples of C/EBPβ ΔuORF mice	RNA-seq of coding RNA	Mus musculus	24	16/01/2019	-		
<a href="#">E-MTAB-7020</a>	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	-		-
<a href="#">E-MTAB-7017</a>	RNASeq data analysis of wild type and reverb 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	-		
<a href="#">E-MTAB-2351</a>	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

Structured metadata is key

- Controlled vocabularies / Ontologies



- The FAIR principles relies on **good data management practices** in all phases of research
  - Research documentation
  - Data organisation
  - Information security
  - Ethics and legislation
- **FAIR data ≠ Open data**  
 Data can be Open without being FAIR  
 Data can be FAIR without being open  
*“As open as possible, as closed as necessary”*





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

# What to do?

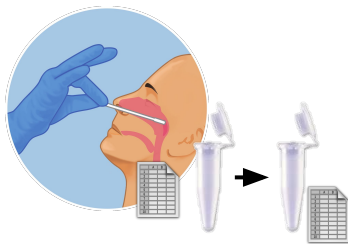
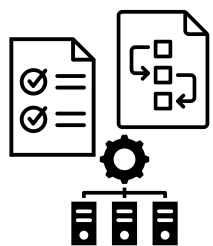
Deposit the data in a data repository!



- What **types** of data will you generate?
- Identify **repositories** to deposit the data in
  - **F1. (meta)data are assigned a globally unique and persistent identifier**
  - *F2. data are described with rich metadata (defined by R1)] - next slide*
  - **F3. metadata clearly and explicitly include the identifier of the data it describes**
  - **F4. (meta)data are registered or indexed in a searchable resource**
  - **A1. (meta)data are retrievable by their identifier using a standardized communication protocol**
    - **A1.1 the protocol is open, free, and universally implementable**
    - *A1.2 the protocol allows for an authentication and authorization procedure, where necessary [controlled access repositories]*
  - *A2. metadata are accessible, even when the data are no longer available*

- Documentation requirements and guidelines
  - Description of study
  - Description of source samples
  - Description of technical treatment (lab methods, instruments, etc)
  - *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*
  - **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**





Study & data  
design

Sampling  
& specimen  
collection

Sample  
preparation

Sample analysis  
& data generation

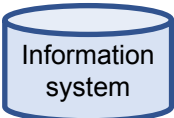
Data processing  
to prepare inputs  
for analysis

Data  
analysis

Communicating  
results

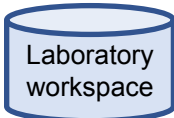
## Procedures

data protection,  
ethics permit,  
infrastructure,  
standards,  
protocols,  
data dictionaries,  
data access, ...



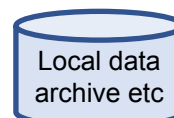
## Biosamples and instruments

populations (statistical) and inclusion criteria,  
physical processing steps,  
working storage conditions,  
long-term storage location,  
sample quality assessment,  
sample annotations,  
reagents, ...



## Data and computational workflows

digital processing steps,  
working storage conditions,  
long-term storage location,  
data quality assessment,  
sample/data annotations,  
reference data, ...



## Outputs

publications,  
data,  
tools,  
workflows,  
reports,  
dashboards, ...

- 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**
- Some motivating factors
  - Democratic principles
  - Good research practices
  - Societal and academic impact

Swedish Research Bill 2021–2024\*

“ [...] *research data shall be made accessible as **open as possible and as closed as necessary***”

\* Our translation from Swedish

The EU's Open Science policy

“ **FAIR [...] open data sharing should become the default for the results of EU-funded scientific research**”



EUROPEAN OPEN  
SCIENCE CLOUD



SND  
Swedish National Data Service



- [Directive \(EU\) 2019/1024](#) of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information
- To be implemented into national member state laws

*"EU countries must adopt policies and take action to make **publicly funded research data openly available**, following the principle of ‘**open by default**’ and support the dissemination of research data that are findable, accessible, interoperable and reusable (the ‘**FAIR**’ principles)"*

# Funders

- Data Management Plans
- Open Data

Vetenskapsrådet, FORMAS, Riksbankens Jubileumsfond

**SUHF**

REK 2019:3  
Dnr. 0005-17  
2019-12-18

**Rekommendation avseende styrdokument för forskningsdata**

SUHF:s styrelse antog den 18 december 2019 följande rekommendation avseende styrdokument för forskningsdata. Rekommendationen har innan beslut varit på remiss hos SUHF:s medlemsinstitutioner.

**Inledning**

Att öppet tillgängliggöra forskningsdata är värdefullt för att validera forskningsresultat och för att möjliggöra återanvändning av forskning i syfte att skapa ny kunskap.

En utveckling både nationellt och internationellt är att fler och fler aktörer strävar efter och ställer krav på att forskningsdata ska vara öppna och tillgängliga. Vissa lösningar behövs till följd av detta i ställning till i vilken utsträckning och på vilket sätt forskningsdata ska göras tillgängliga. Det är också nödvändigt att tydliggöra hanteringen av forskningsdata generellt, dels för att kunna tillgängliggöra den i större utsträckning, dels för att uppfylla befintliga lagar.

SUHF stöder principen att forskning som är helt eller delvis offentligt finansierad ska vara öppet tillgänglig, enligt internationellt vedertagna principer, i den mån det är möjligt med hänsyn till juridiska, etiska och eventuella kommersiella aspekter. Svenska lärostaten bör därför verka för att främja tillgängliggörande av forskningsdata genom forsknings- och utbildningsmiljöer som stödjer, uppmuntrar och informerar om öppen vetenskap som praxis.

Rekommendationen har tagits fram av SUHF:s nationella arbetsgrupp för forskningsdata.

**Guidelines on managing research data**

This document is a translation. In case of a discrepancy between English version of the decision, the Swedish original will prevail.

**Guidelines on managing research data**

This regulation has been approved by the President (ref. V-2019 January 2021). The policy document regulates research data in international principles of good research practice and in accordance with various kinds of data in the field of research. KTH Library is responsible for questions about the policy document.

**1 Managing research data**

KTH strives to ensure that the results of KTH's research shall be accessible to researchers and society. Research data that has been fully funded and forms the basis of published results must therefore be made accessible with due reference to legal, ethical and possible commercial aspects. Internationally adopted FAIR principles to ensure that research data is accessible, interoperable and reusable. These principles are as follows, as clarified as necessary:

**1.1 Creating a data management plan**

Researchers must create a data management plan at an early stage in order to facilitate the management of research data. Such a plan should include information on how the data will be managed, taking into account ethical aspects and in accordance with current legislation. The methods selected must satisfy the requirements of applicable legislation and be clarified whether any stage of the research process involves the creation of research data.

• are a public document in accordance with the Swedish Freedom of the Press Act (1949:105) and how this information is to be made available;

• are a general document subject to confidentiality in accordance with the Swedish Public Access to Information and Secrecy Act (2009:400) and that management takes place to ensure that confidentiality is not prejudiced;

**Stockholms universitet**

Rektor

Handg.: Wilhelm Widmark  
Överbibliotekarie

BESLUT  
2018-02-22

Der SU FV-S.1.1-1780-17

**Forskningsdatapolicy**

Forskning som är helt eller delvis offentligt finansierad bör hanteras och vara öppet tillgänglig enligt internationellt vedertagna principer i den mån det är möjligt med hänsyn till juridiska, etiska och eventuella kommersiella aspekter. Stockholms universitet stödjer de internationella FAIR-dataprinciperna som innebär att forskningsdata ska hanteras på ett sätt som gör dem sökbara, tillgängliga, interoperabla och återanvändningsbara. Att öppet tillgängliggöra forskningsdata eller information om data är värdelöst för att validera forskningsresultat och möjliggöra återanvändning av forskning för att skapa ny kunskap.

Version 20210107

UPV 2021/28

**Riktlinjer för hantering av forskningsdata vid Uppsala universitet**

Ämnet om riktlinjerna

Riktlinjer beskriver Uppsala universitetets principer för hantering av forskningsdata som sammanfattas och används av lärostatens forskare i ett vetenskapligt syfte. Riktlinjerna gäller såväl forskare som forskare vid universitetet som för data som lagras på annan plats. Riktlinjerna är baserade på SUHF:s rekommendation avseende styrdokument för forskningsdata.

Uppsala universitet rekommenderar öppet tillgängliggörande av lärostatens forskningsdata när så är rimligt med hänsyn till juridiska, etiska, tekniska och kommersiella aspekter. Öppet tillgängliggörande innebär att forskningsdata ska vara tillgängliga för alla och att forskarna ska ha rätt att dela sina data. Detta innebär att forskarna ska ha rätt att dela sina data och att forskarna ska ha rätt att dela sina data. Detta innebär att forskarna ska ha rätt att dela sina data och att forskarna ska ha rätt att dela sina data.

Uppsala universitet rekommenderar att forskningsdata ska hanteras på ett sätt som gör dem sökbara, tillgängliga, interoperabla och återanvändningsbara. Detta innebär att forskarna ska ha rätt att dela sina data och att forskarna ska ha rätt att dela sina data. Detta innebär att forskarna ska ha rätt att dela sina data och att forskarna ska ha rätt att dela sina data.

**Riktlinjer för forskningsdokumentation och datahantering vid Karolinska Institutet**

Dnr 1-20/2021

Gäller från och med 2021-01-26

# Universities Research Data Policies



**How do you currently support NGI users to make their data FAIR?**

