NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

SciLifeLab

Open Science & FAIR

Introduction to Data Management Practices course

NBIS DM Team

data@nbis.se

https://nbisweden.github.io/module-open-science-dm-practices/index.html

Open Science

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

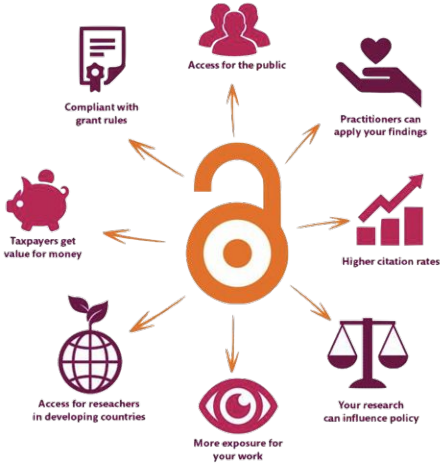
- Open methodology
- **Open source**
- **Open data**
- Open access
- Open peer review
- Open educational resources

Reasons for Open Science

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


Access for the public
Compliant with grant rules
Practitioners can apply your findings
Taxpayers get value for money
Higher citation rates
Access for reseachers in developing countries
More exposure for your work
Your research can influence policy

Picture source: Karolinska institute library

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

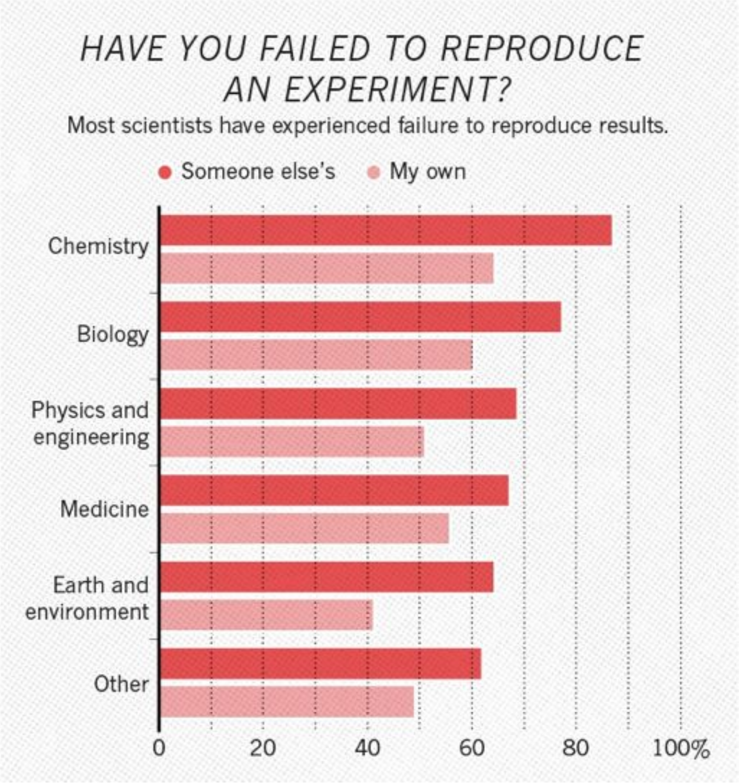
My take of material by Rochelle Tractenberg "Unexpected Ethical Challenges in Bioinformatics and Genomics."

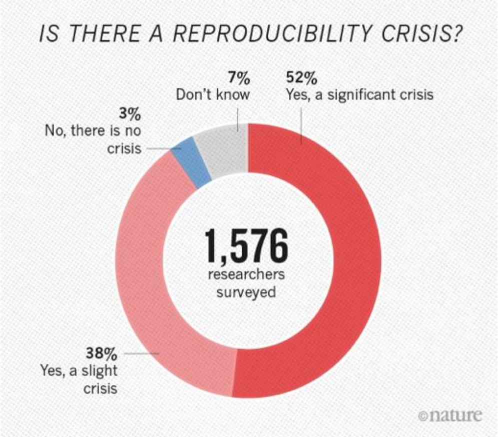
Crisis?

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

If so, what are some of its causes?

A reproducibility crisis



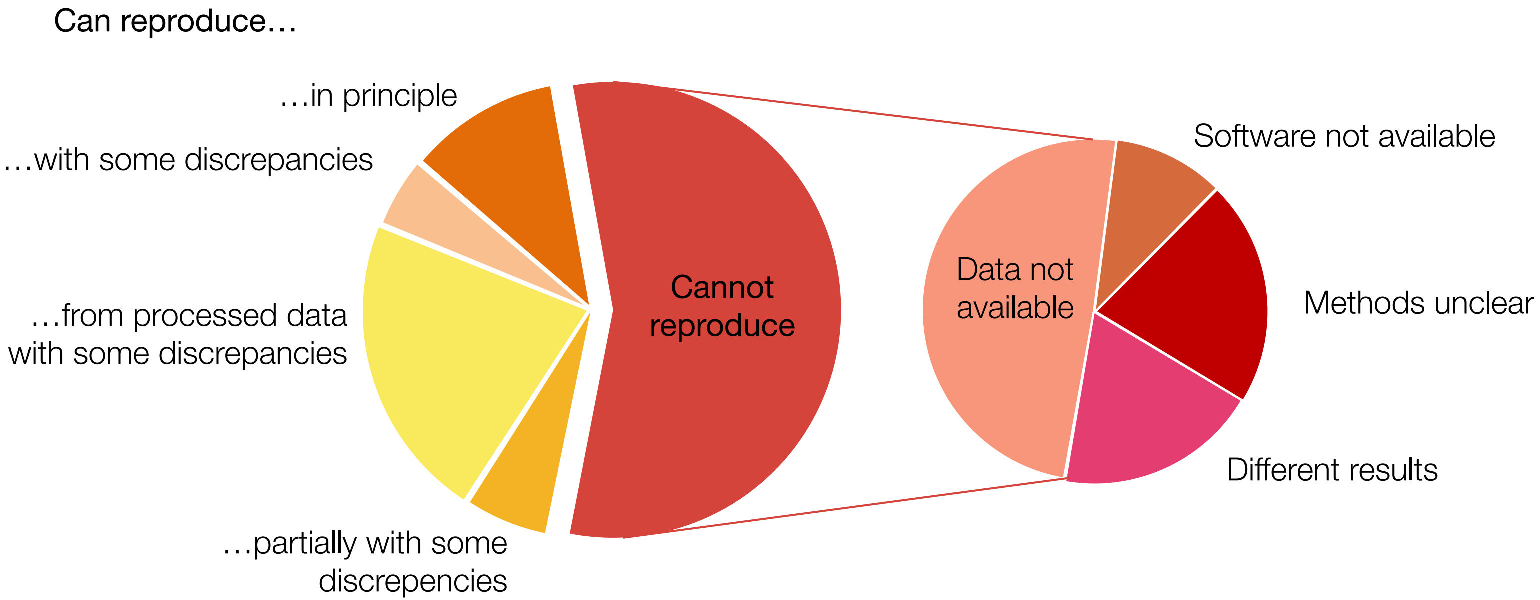


[1] "1,500 scientists lift the lid on reproducibility". Nature. 533: 452–454

[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:



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

Hello! My name is Dr. Judy Benign, I'm an oncologist at NYU School of Medicine.

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

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

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

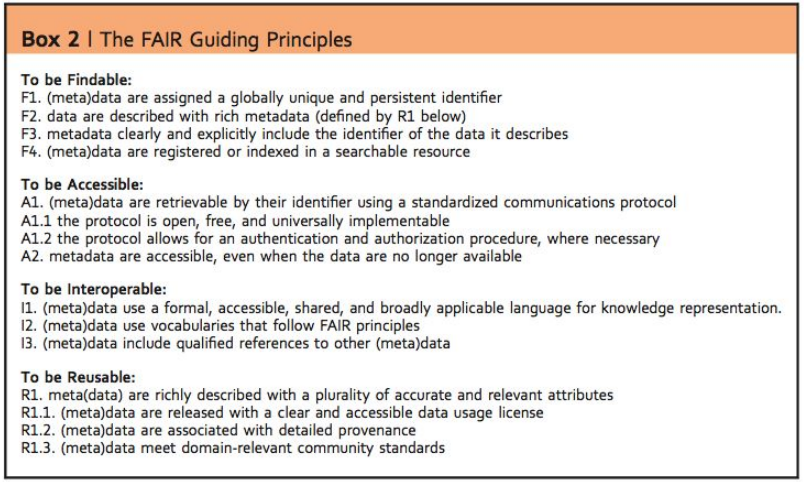
Comment: The FAIR Guiding Principles for scientific data management and stewardship

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 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



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

When to be FAIR?

FAIR at source?

Research happens…

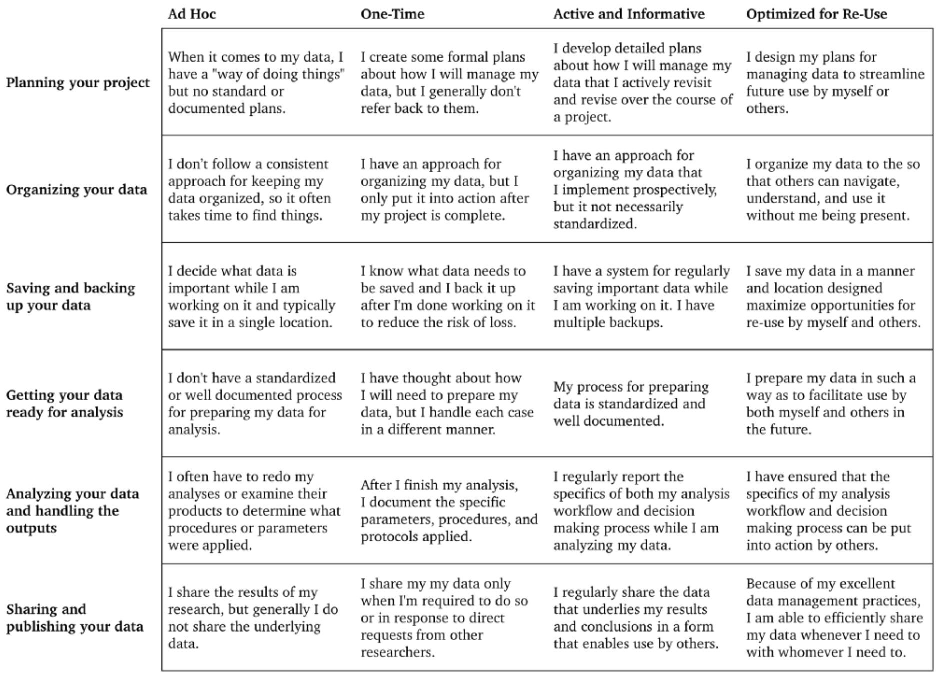
© rassco

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

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**
- 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

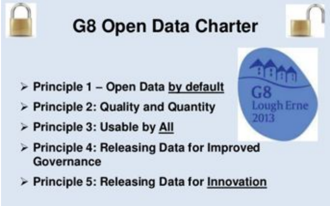
elixir SWEDEN

SND Swedish National Data Service

The Political landscape

OPEN ACCESS

- Strong international movement towards Open Science
- European Commission recommended the member states to establish national guidelines for Open Access
  - Swedish Research Council (VR) submitted proposal to the government Jan 2015
- Research bill 2017–2020 – *28 Nov 2016*
  - “*The aim of the government is that all scientific publications that are the result of publicly funded research should be openly accessible as soon as they are published. Likewise,* ***research data*** *underlying scientific publications should be* ***openly accessible*** *at the time of publication.*” [my translation]
- 2018 – VR assigned by the government to coordinate national efforts to implement open access to research data



G8 Open Data Charter

G8
Lough Erne
2013

- Principle 1 – Open Data by default
- Principle 2: Quality and Quantity
- Principle 3: Usable by All
- Principle 4: Releasing Data for Improved Governance
- Principle 5: Releasing Data for Innovation

Vetenskapsrådet

FÖRSLAG TILL NATIONELLA RIKTLINJER FÖR ÖPPEN TILLGÅNG TILL VETENSKAPLIG INFORMATION

VETENSKAPSRÅDETS RAPPORTER

Regeringens proposition 2016/17:50

Kunskap i samverkan – för samhällets utmaningar och stärkt konkurrenskraft

Prop. 2016/17:50

Regeringen överlämnar denna proposition till riksdagen.

Stockholm den 24 november 2016

Stefan Löfven

Helene Hellmark Knutsson
(Utbildningsdepartementet)

Propositionens huvudsakliga innehåll

I propositionen presenterar regeringen sin syn på forskningspolitikens inriktning i ett tioårigt perspektiv, med särskilt fokus på satsningar 2017–2020. Syftet är att Sverige ska vara en ledande kunskapsnation och ett av världens främsta forsknings- och innovationsländer.

En utgångspunkt är att värna den fria forskningen samtidigt som forskningspolitiken svarar mot globala och nationella samhällsutmaningar. Prioriterade utmaningar är klimat och miljö, hälsa, ökad digitalisering, ett hållbart samhälle och förbättrade kunskapsresultat i det svenska skol- och utbildningssystemet. Ett antal satsningar som stärker den fria forskningen och svarar mot samhällsutmaningarna presenteras.

Förutsättningarna för universitetens och högskolornas långsiktiga strategiska planering bör stärkas genom ökade anslag för forskning och utbildning på forskarnivå. Även lärosätenas samverkan med det omgivande samhället och sambandet mellan utbildning och forskning förväntas stärkas. Vidare redovisar regeringen ett antal åtgärder för att främja attraktiva villkor för studerande på forskarnivå och unga forskare och för att främja jämställdhet i högskola och forskning.

Regeringen har i budgetpropositionen för 2017 lämnat förslag och redovisat bedömningar om fördelning av medel till forskning och innovation. I denna proposition beskrivs satsningarna närmare. Forskningssatsningarna avser dels långsiktiga nationella forskningsprogram som utgår från de utpekade samhällsutmaningarna, dels vissa särskilda forskningssatsningar.

Satsningarna på innovation avser bl.a. en förstärkning av strategiska innovationsområden, vilka ska kopplas till prioriteringarna i regeringens

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)’

"G20 Summit in Hangzhou" by Organisation for Economic Co-operation and Develop is licensed under CC BY-NC 2.0

EOSC

EUROPEAN OPEN SCIENCE CLOUD

The EOSC will offer 1.7 million European **researchers** and 70 million professionals in science, technology, the humanities and social sciences a virtual environment with **open and seamless services for storage, management, analysis and re-use of research data**, across borders and scientific disciplines by federating existing scientific data infrastructures, currently dispersed across disciplines and the EU Member States.

EOSCFAIR
Executive Board Working Group

https://www.eosc-portal.eu/

“Open Data Directive”

- 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

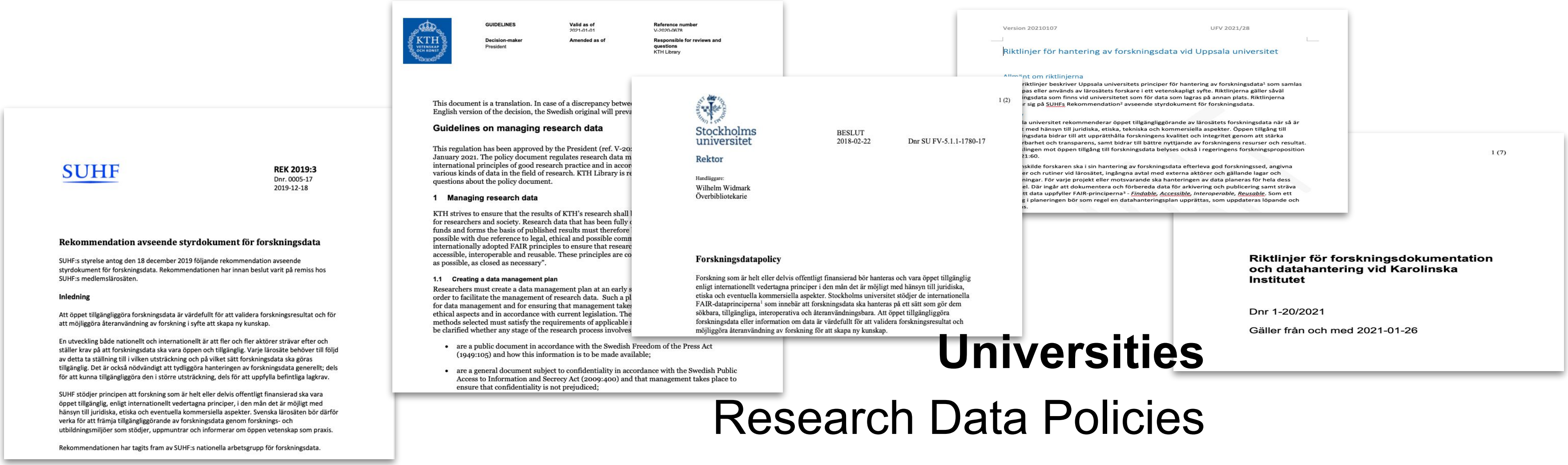
"EU countries must adopt policies and take action to make ***publicly funded research data openly available****, following the principle of ‘****open by default****’ and support the dissemination of research data that are findable, accessible, interoperable and reusable (the ‘****FAIR’ principles****)"*

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 medlemslärosäten.

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 öppen och tillgänglig. Varje lärosäte behöver till följd av detta ta ställning till i vilken utsträckning och på vilket sätt forskningsdata ska göras tillgänglig. Det är också nödvändigt att tydliggöra hanteringen av forskningsdata generellt; dels för att kunna tillgängliggöra den i större utsträckning, dels för att uppfylla befintliga lagkrav.

SUHF stödjer 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 lärosäten 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.

KTH VETENSKAP OCH KONST

GUIDELINES	Valid as of 2021-01-01	Reference number V-2020-0678
Decision-maker President	Amended as of	Responsible for reviews and questions KTH Library

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

Guidelines on managing research data

This regulation has been approved by the President (ref. V-20
January 2021. The policy document regulates research data m
international principles of good research practice and in accor
various kinds of data in the field of research. KTH Library is re
questions about the policy document.

1 Managing research data

KTH strives to ensure that the results of KTH's research shall
for researchers and society. Research data that has been fully
funds and forms the basis of published results must therefore
possible with due reference to legal, ethical and possible comm
internationally adopted FAIR principles to ensure that researc
accessible, interoperable and reusable. These principles are co
as possible, as closed as necessary".

1.1 Creating a data management plan

Researchers must create a data management plan at an early s
order to facilitate the management of research data. Such a pl
for data management and for ensuring that management takes
ethical aspects and in accordance with current legislation. The
methods selected must satisfy the requirements of applicable
be clarified whether any stage of the research process involves

- 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

1 (2)

BESLUT
2018-02-22

Dnr SU FV-5.1.1-1780-17

Handläggare:
Wilhelm Widmark
Överbibliotekarie

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[1] som innebär att forskningsdata ska hanteras på ett sätt som gör dem sökbara, tillgängliga, interoperativa och återanvändningsbara. Att öppet tillgängliggöra forskningsdata eller information om data är värdefullt för att validera forskningsresultat och möjliggöra återanvändning av forskning för att skapa ny kunskap.

Version 20210107

UFV 2021/28

Riktlinjer för hantering av forskningsdata vid Uppsala universitet

Allmänt om riktlinjerna

riktlinjer beskriver Uppsala universitets principer för hantering av forskningsdata[1] som samlas
pas eller används av lärosätets forskare i ett vetenskapligt syfte. Riktlinjerna gäller såväl
ingsdata som finns vid universitetet som för data som lagras på annan plats. Riktlinjerna
r sig på SUHFs Rekommendation[2] avseende styrdokument för forskningsdata.

la universitet rekommenderar öppet tillgängliggörande av lärosätets forskningsdata när så är
t med hänsyn till juridiska, etiska, tekniska och kommersiella aspekter. Öppen tillgång till
ingsdata bidrar till att upprätthålla forskningens kvalitet och integritet genom att stärka
rbarhet och transparens, samt bidrar till bättre nyttjande av forskningens resurser och resultat.
lingen mot öppen tillgång till forskningsdata belyses också i regeringens forskningsproposition
21:60.

nskilde forskaren ska i sin hantering av forskningsdata efterleva god forskningssed, angivna
er och rutiner vid lärosätet, ingångna avtal med externa aktörer och gällande lagar och
ningar. För varje projekt eller motsvarande ska hanteringen av data planeras för hela dess
el. Där ingår att dokumentera och förbereda data för arkivering och publicering samt sträva
tt data uppfyller FAIR-principerna[3] - *Findable, Accessible, Interoperable, Reusable*. Som ett
g i planeringen bör som regel en datahanteringsplan upprättas, som uppdateras löpande och
ıs.

1 (7)

Riktlinjer för forskningsdokumentation och datahantering vid Karolinska Institutet

Dnr 1-20/2021

Gäller från och med 2021-01-26

Universities

Research Data Policies

Motivators

https://www.pngwing.com/en/free-png-nlacv

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