

UNIVERSITY OF OSLO

Sharing and archiving

research data

Agata Bochynska and Live Håndlykken Kvale
Open Research, University of Oslo Library
CC-BY-SA-4.0 2022



Agenda

- **Why archive research data?**
 - Requirements and new assessment criteria
 - Reasons for sharing
 - How to share data?
- **Data repositories**
 - Types of repositories
 - Finding and selecting repositories
 - Archiving of code
 - Data papers
 - Selecting and preparing data
- **How open?**
 - Levels of FAIR
 - Licenses
- **Q&A**
- **Examples**



“**Open Science** is becoming the modus operandi for carrying out research and innovation by **sharing knowledge, data and tools** as early as possible, in open collaboration with all relevant knowledge actors and society.”

Open science means transparency and **knowledge-sharing** in research processes to make knowledge **accessible** across academic groups, sectors and national boundaries. The concept of open science encompasses the entire research process [...].

- The Research Council of Norway. Policy for open science 2020

The Research Council Policy for Open Science

In effect from 2020



● PLANNED

Researcher Project for Young Talents (Thematic Priority Call)

PUBLISHED 12 OCT 2022 | LAST UPDATED 11 NOV 2022

 SHARE |  DOWNLOAD

Assessment criteria

Communication and exploitation

- **The extent to which the appropriate open science practices are implemented** as an integral part of the proposed project to ensure open sharing and wide distribution of research outputs.
- Quality and scope of communication and engagement activities with different target audiences, including relevant stakeholders/users.

Changing research assessment system



Towards a reform of the research assessment system

Scoping Report

Read final agreement here: https://research-and-innovation.ec.europa.eu/news/all-research-and-innovation-news/reforming-research-assessment-agreement-now-final-2022-07-20_en

<https://op.europa.eu/en/publication-detail/-/publication/36ebb96c-50c5-11ec-91ac-01aa75ed71a1/language-en/>

Changing research assessment system – in Norway



**NOR-CAM - A toolbox
for recognition and rewards
in academic careers**

U:R Universities
Norway

Local Requirements

Research data at the University of Oslo shall:

- be made openly available for further usage
- be made available at an early stage
- have a data management plan
- have metadata and be documented
- must be securely archived
- have licenses for access, reuse and redistribution
- made freely available
- (but the actual distribution cost should be covered)



Source: <https://www.uio.no/english/for-employees/support/research/research-data-management/policies-and-guidelines/>

Local Requirements

Research data at the University of Oslo shall:

be made openly available for further usage

be made available at an early stage

have a data management plan

have metadata and be documented

must be securely archived

have licenses for access, reuse and redistribution

made freely available

(but the actual distribution cost should be covered)



Source: <https://www.uio.no/english/for-employees/support/research/research-data-management/policies-and-guidelines/>

Reasons for sharing your data

External Factors

- Funder and publisher requirements
- Institutional requirements
- New assessment systems

Career Benefits

- Increased visibility
- More data reuse
- New collaborations
- Increased citations

Scientific Progress

- More robust research
- Enables verification of results
- Enables new collaborations across disciplines and borders
- Opens up for new uses of data
- Avoids duplication
- Easier to use data in teaching

How do I share
data?



MadScientist
@MadS100tist



"Data will be available upon request"



Data availability statements don't work

The Availability of Research Data Declines Rapidly with Article Age

Timothy H. Vines,^{1,2,*} Arianne Y.K. Albert,³ Rose L. Andrew,¹ Florence Débarre,^{1,4} Dan G. Bock,¹ Michelle T. Franklin,^{1,5} Kimberly J. Gilbert,¹ Jean-Sébastien Moore,^{1,6} Sébastien Renaut,¹ and Diana J. Rennison¹

¹Biodiversity Research Centre, University of British Columbia, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada

²Molecular Ecology Editorial Office, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada

³Women's Health Research Institute, 4500 Oak Street, Vancouver, BC V6H 3N1, Canada

⁴Centre for Ecology & Conservation Biosciences, University of Exeter, Cornwall Campus, Tremough, Penryn TR10 9EZ, UK

⁵Institute for Sustainable Horticulture, Kwantlen Polytechnic University, 12666 72nd Avenue, Surrey, BC V3W 2M8, Canada

⁶Department of Biology, Université Laval, 1030 Avenue de la Médecine, Laval, QC G1V 0A6, Canada

Summary

Policies ensuring that research data are available on public archives are increasingly being implemented at the government [1], funding agency [2–4], and journal [5, 6] level. These policies are predicated on the idea that authors are poor stewards of their data, particularly over the long term [7], and indeed many studies have found that authors are often unable or unwilling to share their data [8–11]. However, there are no systematic estimates of how the availability of research data changes with time since publication. We therefore requested data sets from a relatively homogenous set of 516 articles published between 2 and 22 years ago, and found that availability of the data was strongly affected by

sets (23%) were confirmed as extant. [Table 1](#) provides a breakdown of the data by year.

We used logistic regression to formally investigate the relationships between the age of the paper and (1) the probability that at least one e-mail appeared to work (i.e., did not generate an error message), (2) the conditional probability of a response given that at least one e-mail appeared to work, (3) the conditional probability of getting a response that indicated the status of the data (data lost, data exist but unwilling to share, or data shared) given that a response was received, and, finally, (4) the conditional probability that the data were extant (either “shared” or “exists but unwilling to share”) given that an informative response was received.

There was a negative relationship between the age of the paper and the probability of finding at least one apparently working e-mail either in the paper or by searching online (odds ratio [OR] = 0.93 [0.90–0.96, 95% confidence interval (CI)], $p < 0.00001$). The odds ratio suggests that for every year since publication, the odds of finding at least one apparently working e-mail decreased by 7% ([Figure 1A](#)). Since we searched for e-mails in both the paper and online, four factors contribute to the probability of finding a working e-mail: (1) the number of e-mails in the paper and (2) the chance that any of those worked and (3) the number of e-mails we could find by searching online and (4) the chance that any of those worked. The total number of e-mail addresses we found in the paper decreased with age (Poisson regression coefficient = -0.07 , $SE = 0.01$, $p < 0.0001$) from an average of 1.17 in 2011 to 0.42 in 1991 ([Figure 2A](#)), and there was a slight positive effect of article age on the number of e-mails we found online (Poisson regression coefficient = 0.015 , $SE = 0.007$, $p < 0.05$; [Figure 2C](#)). Moreover, the chance that an e-mail found in the

Data availability statements don't work



Dorothy Bishop
@deevybee

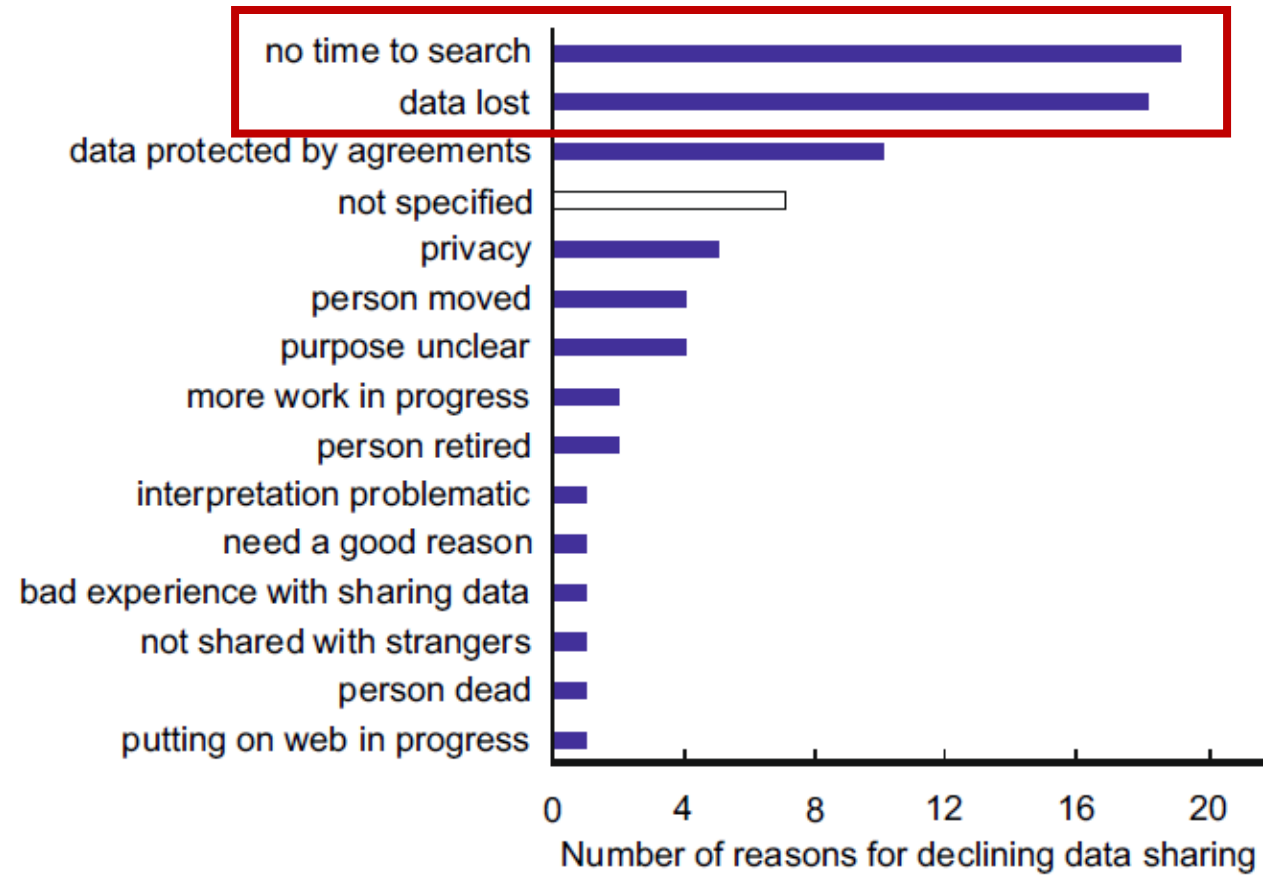


Just had request for data from 2011, which confirms all the advice I've been given:

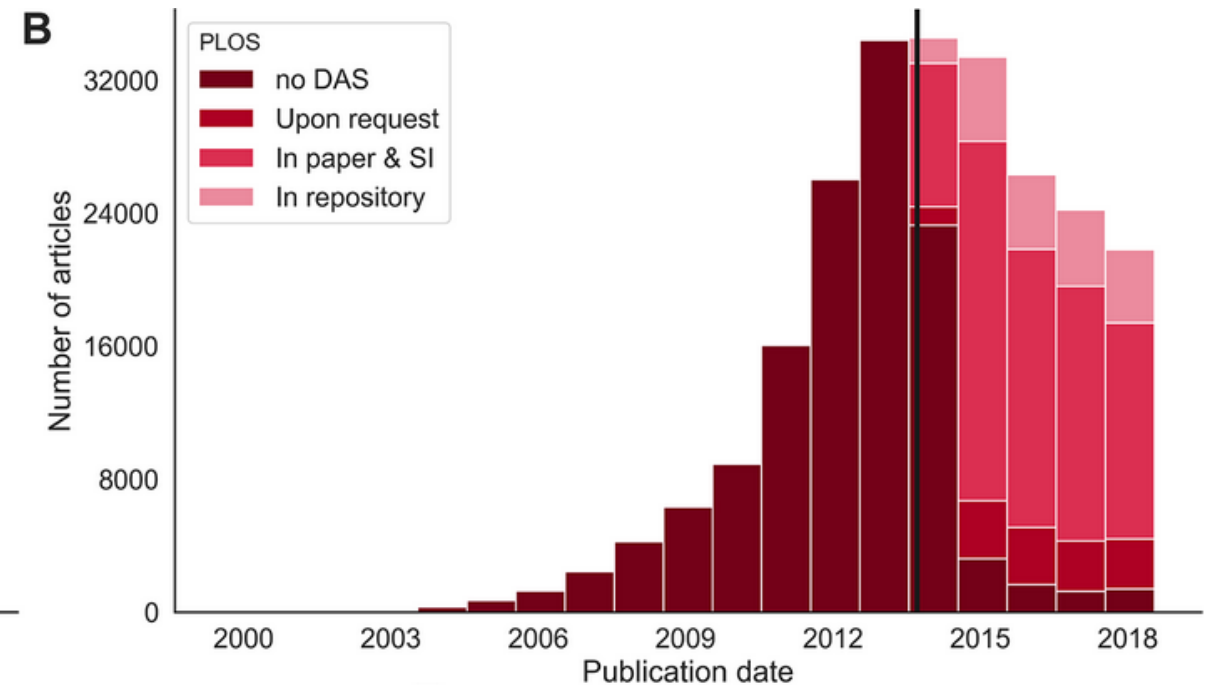
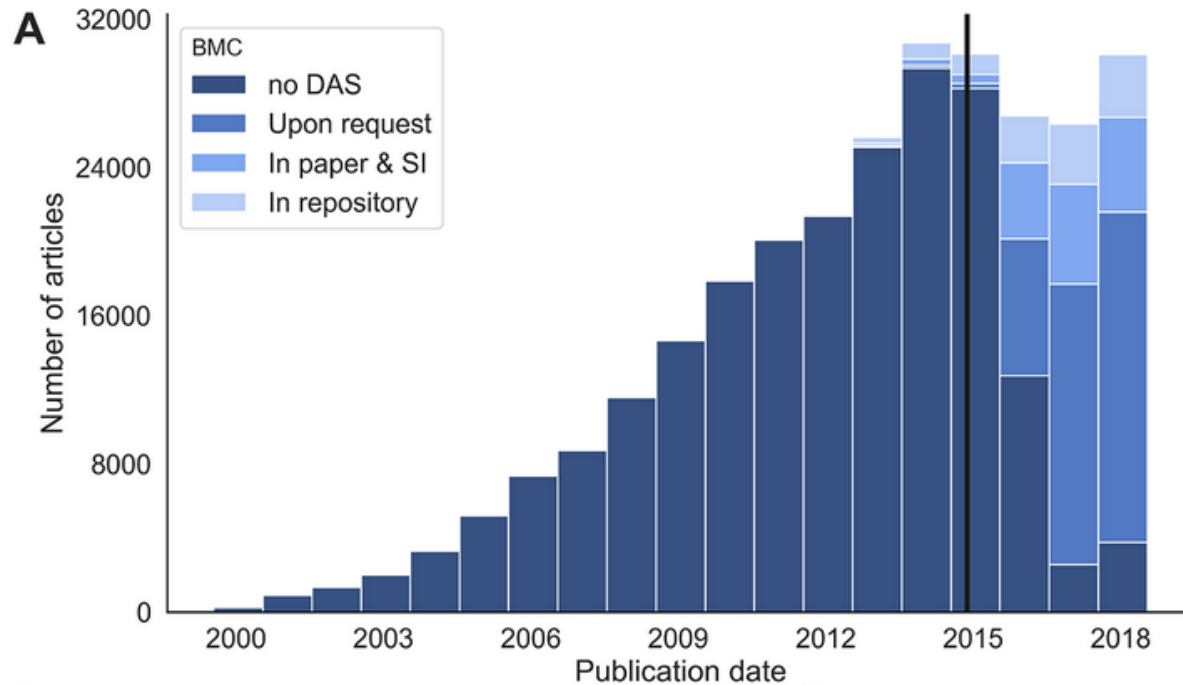
- You will not remember old passwords
- Old hard drives won't work any more
- You will have many files called 'final'
- Keep your scripts as well as data
- Store open data for your future self!

11:45 AM · Sep 22, 2021 · Twitter Web App

Data availability statements don't work



Increasing number of deposited data

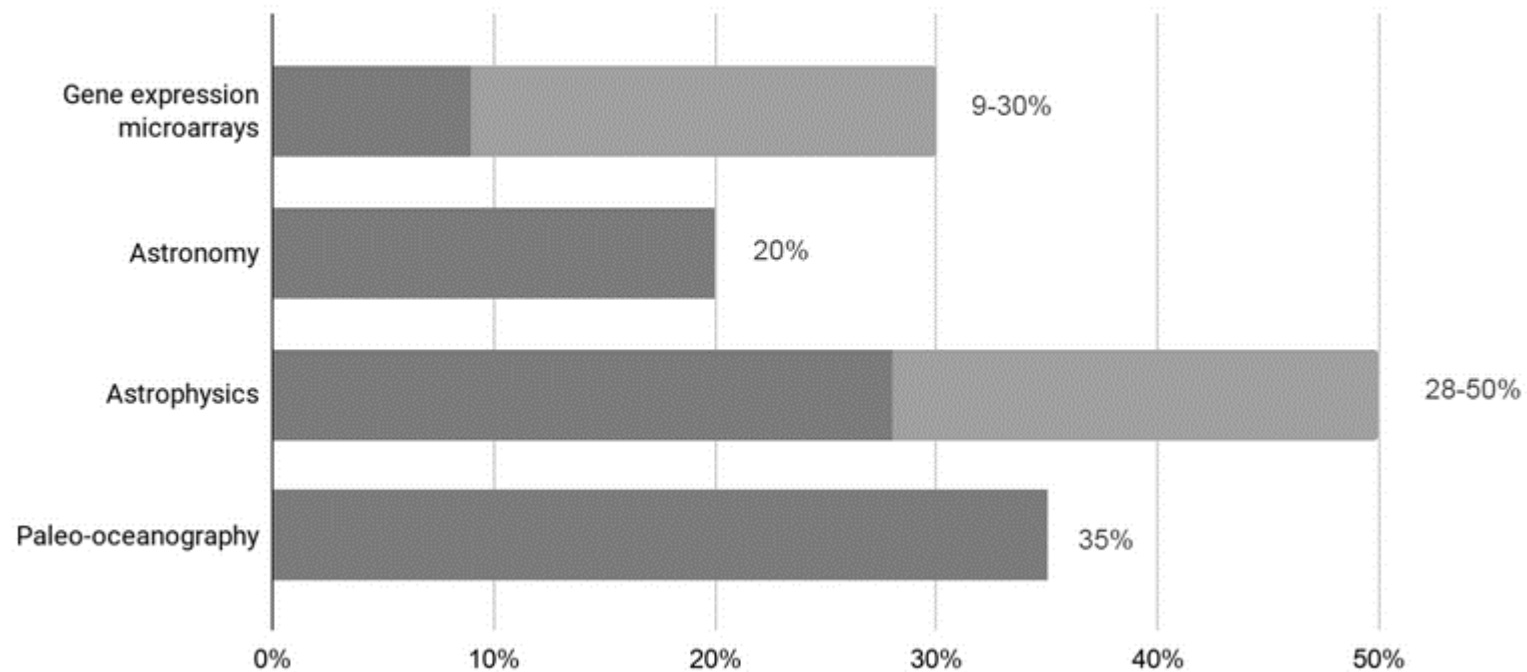


DAS: Data Availability
Statement

Deposited data – more citations

*We also find an association between articles that include statements that link to data in a repository and up to **25.36% ($\pm 1.07\%$) higher citation impact** on average, using a citation prediction model.*

Citation impact of sharing data, by discipline



Deposited data – more citations

*We also find an association between articles that include statements that **link to data in a repository** and up to **25.36% ($\pm 1.07\%$) higher citation impact** on average, using a citation prediction model.*

YES, a repository is the
best place to archive data

Domain-specific data repositories



General-purpose data repositories



Open Science Framework



National or institutional data repositories



NIRD RESEARCH DATA ARCHIVE

<https://dataverse.org/>

<https://www.nsd.no/>

<https://www.sigma2.no/nird-service-platform>

UiO's institutional data archive





UNIVERSITY OF OSLO

DataverseNO >

[✉ Contact](#) [↗ Share](#)



[Advanced Search](#)

☒ [Dataverses \(0\)](#)

☒ [Datasets \(21\)](#)

☐ [Files \(437\)](#)

Publication Year

2022 (7)

2021 (7)

2020 (5)

2018 (2)

Subject

[Earth and Environmental Sciences \(13\)](#)

[Social Sciences \(6\)](#)

[Physics \(3\)](#)

[Agricultural Sciences \(1\)](#)

[Mathematical Sciences \(1\)](#)

[More...](#)

Keyword Term

[Faults \(7\)](#)

[Svalbard \(7\)](#)

[Billefjorden Group \(3\)](#)

[Devonian \(3\)](#)

[Ebbadalen \(3\)](#)

[More...](#)

1 to 10 of 21 Results

↕ Sort ▾

[Structural field measurements in Proterozoic basement, Devonian, and Carboniferous rocks in Kongsfjorden, September 2022](#) 

Nov 2, 2022



Koehl, Jean-Baptiste P.; Stokmo, Eirik M. B., 2022, "Structural field measurements in Proterozoic basement, Devonian, and Carboniferous rocks in Kongsfjorden, September 2022", <https://doi.org/10.18710/APGAWL>, DataverseNO, V1

Structural field measurements (bedding surfaces, brittle to ductile fault surfaces and associated brittle fault lineations, foliation surfaces, fold axes, fold axial planes) from Proterozoic basement, Devonian, and Carboniferous rocks from September 2022 expedition to Kongsfjorde...

[Replication data for the MULTICLIM project "Pesticide effects on the abundance of springtails and mites in field mesocosms at an agricultural site"](#) 

Nov 2, 2022



Konestabo, Heidi Sjursen, 2022, "Replication data for the MULTICLIM project "Pesticide effects on the abundance of springtails and mites in field mesocosms at an agricultural site"", <https://doi.org/10.18710/QWIDIT>, DataverseNO, V1

This dataset was collected as part of the MULTICLIM project at the University of Oslo: <https://www.mn.uio.no/ibv/english/research/sections/aqua/research-projects/144612/> The data has been published in Ecotoxicology, doi:10.1007/s10646-022-02599-3. The main aim of the study for wh...

[Field photographs Kongsfjorden September 2022](#) 

Oct 26, 2022



Koehl, Jean-Baptiste P.; Stokmo, Eirik M. B., 2022, "Field photographs Kongsfjorden September 2022", <https://doi.org/10.18710/KEB2MM>, DataverseNO, V1

Geological fieldwork photographs of September 2022 excursion to Kongsfjorden. Day 0: flight journey onboard Lufttransport plane from Longyearbyen to Ny-Ålesund. Day 1: trip by zodiac boat from Ny-Ålesund to southern Blomstrandhalvøya; landing site: London. Then walking southeastw...

<https://dataverse.no/dataverse/uiio>

Certification



international, community based, non-governmental, and non-profit organization promoting sustainable and trustworthy data infrastructures

The Qualitative Data Repository

Search data...

Discover Data

Deposit Data



QDR
The Qualitative
Data Repository

More Info

About
Terms & Conditions
Policies
Site Map

Contact Info

Qualitative Data Repository
347 Eggers Hall
Maxwell School, Syracuse University
Syracuse, NY 13244
(315) 443 - 6198
qdr@syr.edu

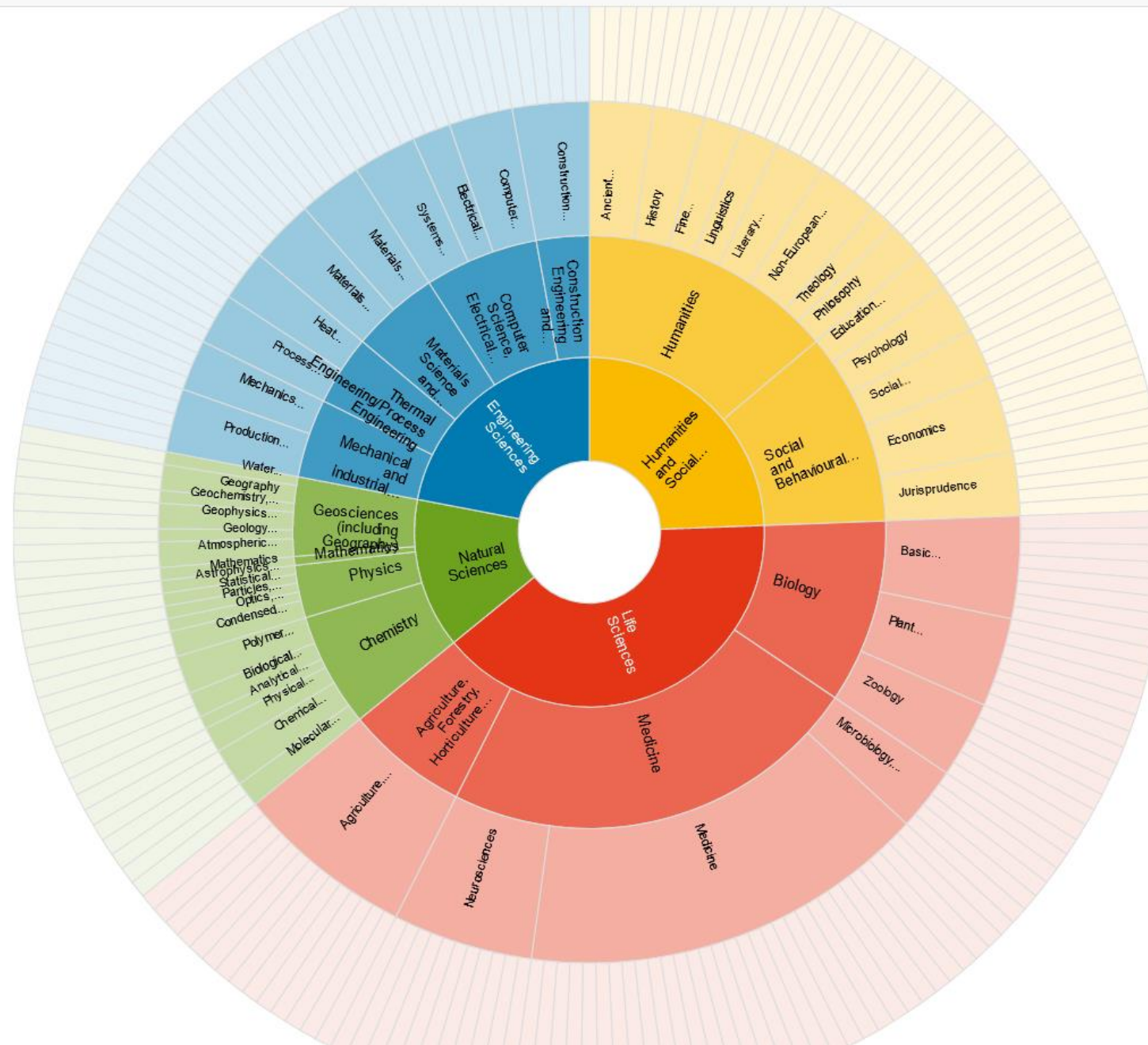
CoreTrustSeal



Certified as compliant with CoreTrustSeal standards.

Finding repositories





Finding repositories

Databases

A registry of knowledgebases and repositories of data and other digital assets.

Search through current results.

SEARCH

MATCH ALL TERMS

MATCH ANY TERM

MAINTAINED

NOT MAINTAINED

RECOMMENDED

NOT RECOMMENDED

READY

DEPRECATED

UNCERTAIN

IN DEV.

Registry

Search

APPLY

Record Type

Search

APPLY

Subjects

Search

APPLY

Domains

Licence(s)

Clear All

Registry: Database

Registry: database

Record Type: repository

Subjects: humanities and social science

<

1

2

3

4

>

Displaying 1 to 30 of 112.

NAHDAP

National Addiction & HIV Data Archive Program

NAHDAP acquires, preserves and disseminates data relevant to drug addiction and HIV research. The scope of the data housed at NAHDAP covers a wide range of legal and illi...

Social Scie...

Drug

Homo sapi...

+4 more tags

Standards Implemented2

Related Databases0

Endorsing Policies11

ICPSR

Inter-university Consortium for Political and Social Research

ICPSR is a data archive of behavioral and social science research data. An international consortium of more than 750 academic institutions and research organizations, ICP...

Economics

Behavior

Homo sapi...

+11 more tags

Standards Implemented5

Related Databases4

Endorsing Policies11

RDA

SND

Selecting an archive to deposit data

- Should the **data** be openly available?
- Should the **metadata** be openly available?
- Presence of **personal or confidential data** can affect choice of archive / preservation solution
- What can the relevant archives offer for **long term** perspectives?
- Does the archive offer **curation** – control of metadata and updating of formats?

Data paper

scientific **data**



Data in Brief
Open access

**Research Data Journal
for the Humanities
and Social Sciences**

<https://www.nature.com/sdata/>
<https://www.sciencedirect.com/journal/data-in-brief>
<https://brill.com/view/journals/rdj/rdj-overview.xml>

nature > scientific data

Reef Cover, a coral reef classification for global habitat mapping from remote sensing

Emma V. Kennedy, Chris M. Roelfsema ... Paul Tudman

Data Descriptor | 02 August 2021



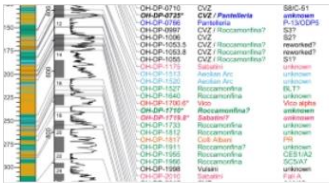
Ishan @seefromthesky via Unsplash

Featured

Data Descriptor
Open Access
02 Sept 2021

Lake Ohrid's tephrochronological dataset reveals 1.36 Ma of Mediterranean explosive volcanic activity

Niklas Leicher, Biagio Giaccio ... Bernd Wagner



16.11.2022

Data Descriptor
Open Access

OPERA tau neutrino charged current interactions



Archiving of Code



Selecting data for archiving

- Does your dataset have a potential for reuse?
- (Inter-)national or historical importance
- Data quality
- Uniqueness or originality
- Size, scale, cost
- Innovativeness

Preparing for archiving

- Consistent, meaningful, and compatible **file naming**
- Choose accessible, patent-free, and open **file formats**



Text documents

• Preferred format(s)

- PDF/A (.pdf)
- ODT (.odt)

• Non-preferred format(s)

- Microsoft Word (.doc)
- Office Open XML (.docx)
- Rich Text File (.rtf)
- PDF other than PDF/A (.pdf)

Plain text

- Unicode text (.txt)

- Non-Unicode text (.txt)

Markup language

- XML (.xml)
- HTML (.html)
- Related files: .css, .xslt, .js, .es

- SGML (.sgml)
- Markdown (.md)

Programming languages

- MATLAB
- NetCDF
- TextFabric

Spreadsheets

- ODS (.ods)
- CSV (.csv)

- Microsoft Excel (.xls)
- Office Open XML Workbook (.xlsx)
- PDF/A (.pdf)

Databases

- SQL (.sql)
- SIARD (.siard)
- CSV (.csv)

- Microsoft Access (.mdb, .accdb)
- dBase (.dbf)
- HDF5 (.hdf5, .he5, .h5)

Statistical data

- SPSS (.dat/.sps)
- STATA (.dat/.DO)
- R

- SPSS Portable (.por)
- SPSS (.sav)
- STATA (.dta)
- SAS (.7dat; .sd2; .tpt)

Preparing for archiving

- Consistent, meaningful, and compatible **file naming**
- Choose accessible, patent-free, and open **file formats**
- Make sure you have the necessary **documentation** (and metadata)

Preparing for archiving

- Consistent, meaningful, and compatible **file naming**
- Choose accessible, patent-free, and open **file formats**
- Make sure you have the necessary **documentation** (and metadata)
- **Reduce complexity** by grouping large groups of similar files in zip bundles to make upload and download easier
- Consider **size limitations** when choosing an archive (e.g. some archives have a limit of 10 to 50 GB per dataset)

As open as possible,
as closed as necessary

F_{indable}



A_{ccessible}



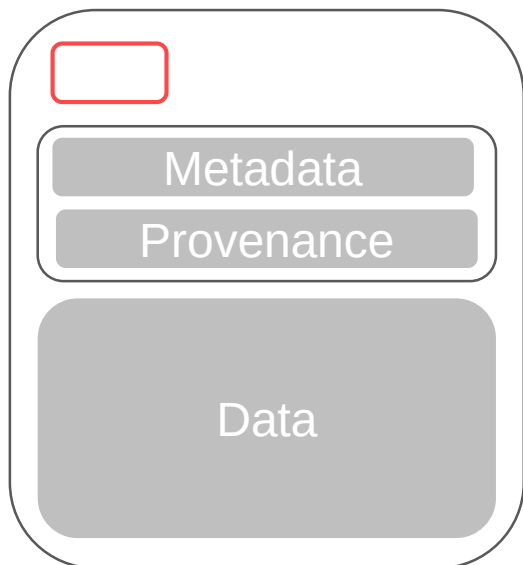
I_{nteroperable}



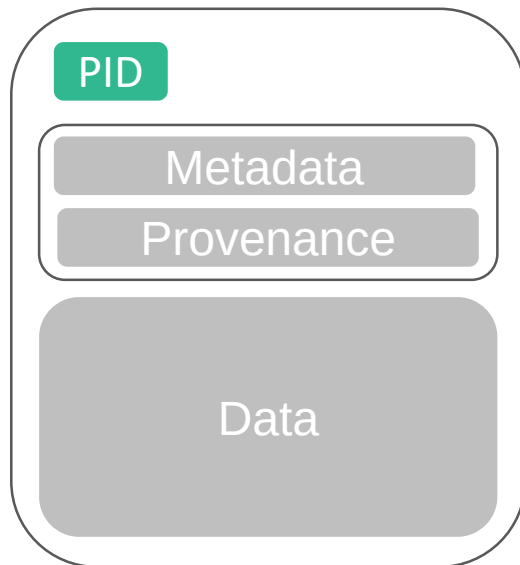
R_{eusable}



Totally UNFAIR



Findable
Usable for humans



Levels of FAIR

Adapted from:
<https://www.force11.org/>

Persistent identifiers (PID)

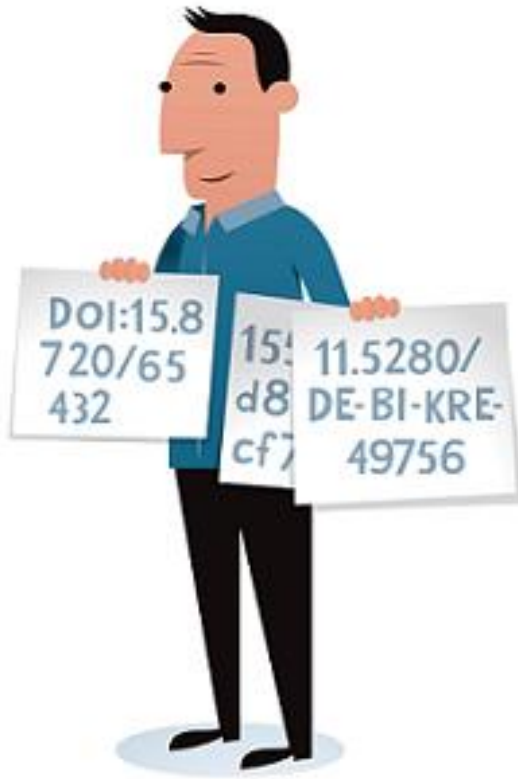
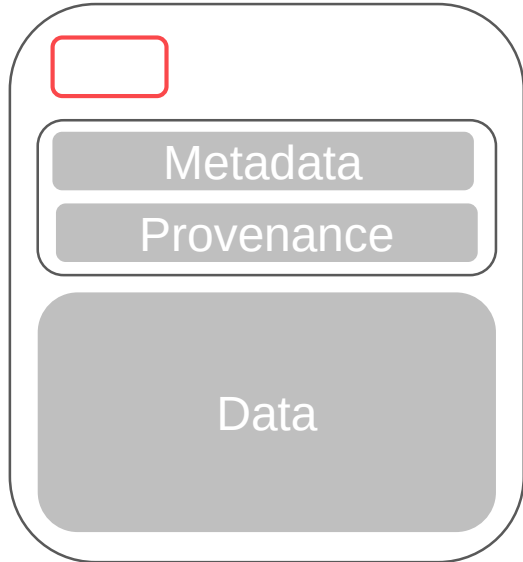


Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

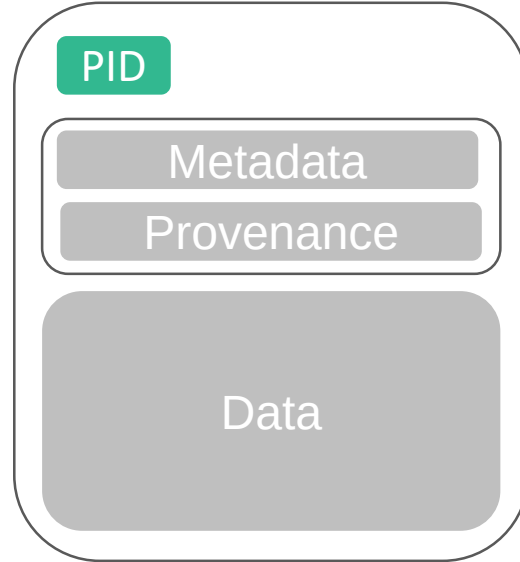
- Published data should have a persistent identifier (PID)
- PID eases referencing a dataset – web url's tend to change or stop working
- Common types of PIDs:
 - **Digital Object Identifier (DOI)**
 - Uniform Resource Name (URN)
 - Permanent Uniform Resource Locator (PURL)

Levels of FAIR

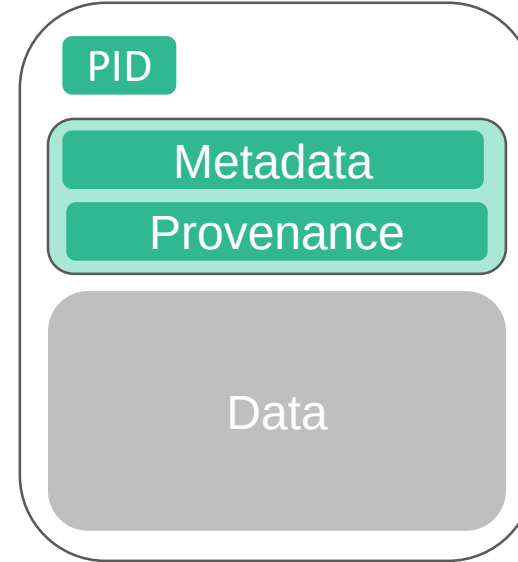
Totally UNFAIR



Findable
Usable for humans

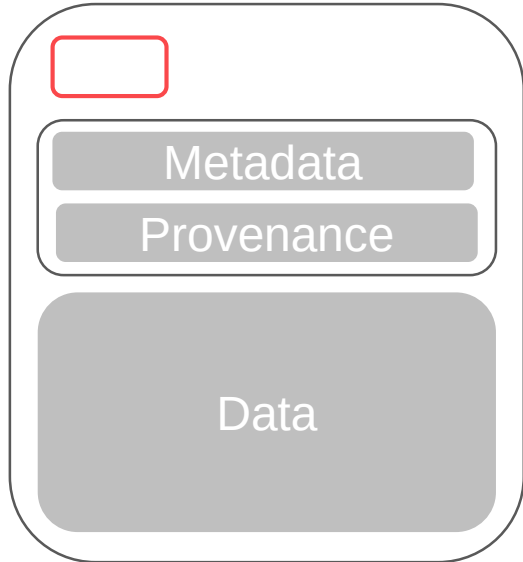


FAIR metadata

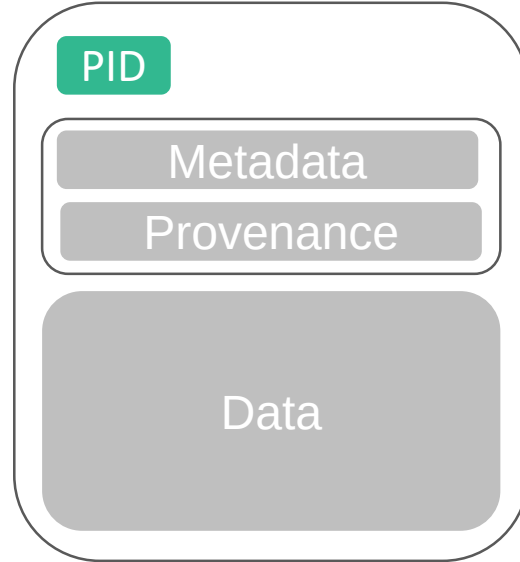


Levels of FAIR

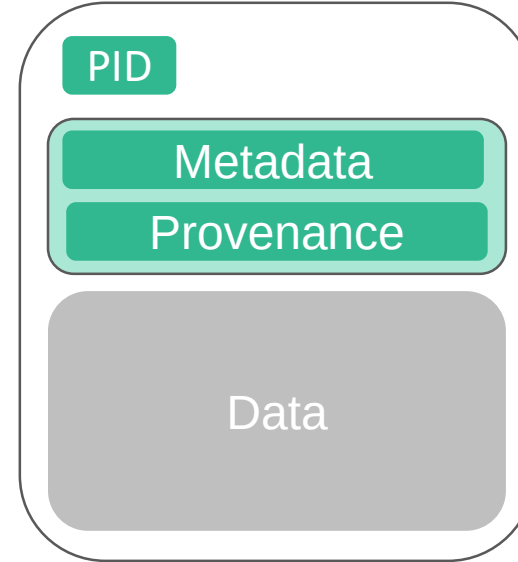
Totally UNFAIR



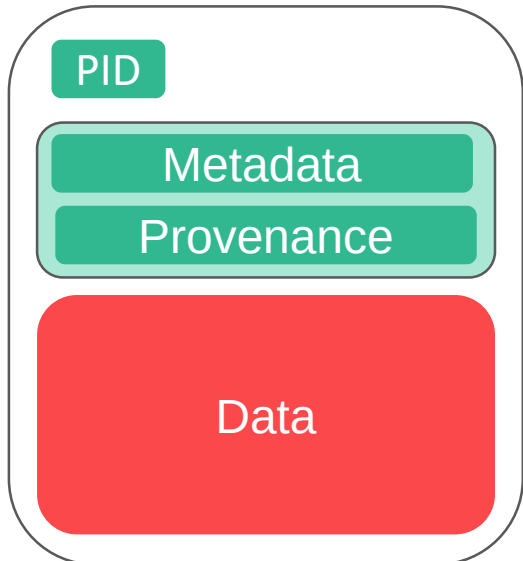
Findable
Usable for humans



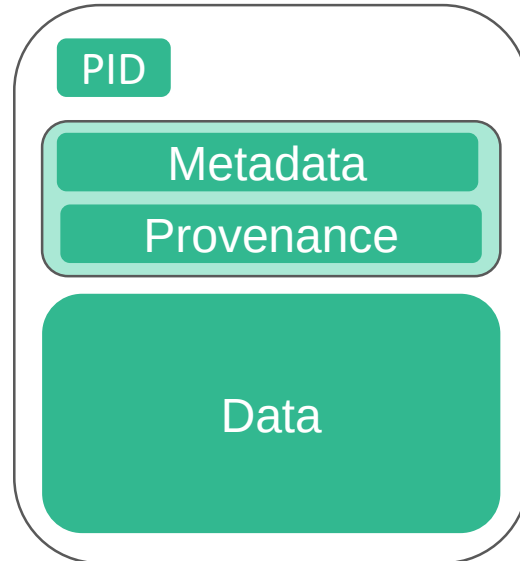
FAIR metadata



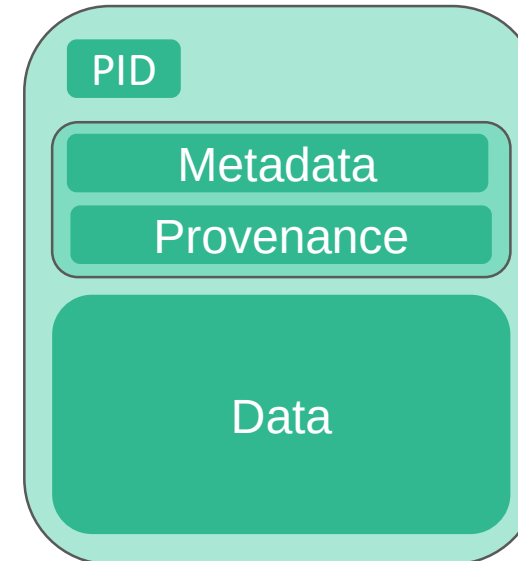
FAIR data
Restricted access



FAIR data
Open access



FAIR data
Open access and functionally linked



As open as possible,
as closed as necessary



- Credit must be given to the creator
- Only noncommercial uses of the work are permitted



- Credit must be given to the creator
- Adaptations must be shared under the same terms



- Credit must be given to the creator



- is a public dedication tool, which allows creators to give up their copyright and put their works into the worldwide public domain.

What is a license?

- A license agreement is a legal arrangement between the creator/depositor of the dataset and the data repository, signifying what a user is allowed to do with the data.
- Creative Commons licenses are often used, more licenses can be found at: <https://spdx.org/licenses/>



License makes data reusable

Things to consider before licensing

- Which rights apply to the dataset (ownership, copyright?)
- Do you hold rights to the dataset?
- If not, who hold rights to the dataset (do you need to ask for necessary permissions?)
- Do others hold rights to the dataset (e.g. when working with secondary/existing data)

Committee on sharing and reuse of research data

Unresolved issues regarding rights and licenses often prevent the results of research from being made publicly available.

The Ministry of Education and Research asked the Research Council of Norway and UNIT to set up a committee to examine rights and licensing issues related to sharing of data sets produced in scientific research. The Research Council served as secretariat for the committee.

Read the final report: [How should we share research data? Report and recommendations relating to licensing and making research data available](https://www.forskningradet.no/en/Adviser-research-policy/open-science/open-access-to-research-data/sharing-reuse-research-data/) (PDF).

Read the report

→ [How should we share research data.pdf](#)

Research data should be assigned a standard international license which:

- Has as few access and use restrictions as possible
- To the largest extent promotes principles of juridical interoperability for research data
- Is human and machine readable

Local Requirements

At the University of Oslo:

Research data shall be provided with licenses for access, reuse and redistribution

- Licenses should be internationally recognized
- Licenses should include as few restrictions as possible in respect of access, reuse, and redistribution of the data



Source: <https://www.uio.no/english/for-employees/support/research/research-data-management/policies-and-guidelines/>



- Credit must be given to the creator
- Only noncommercial uses of the work are permitted



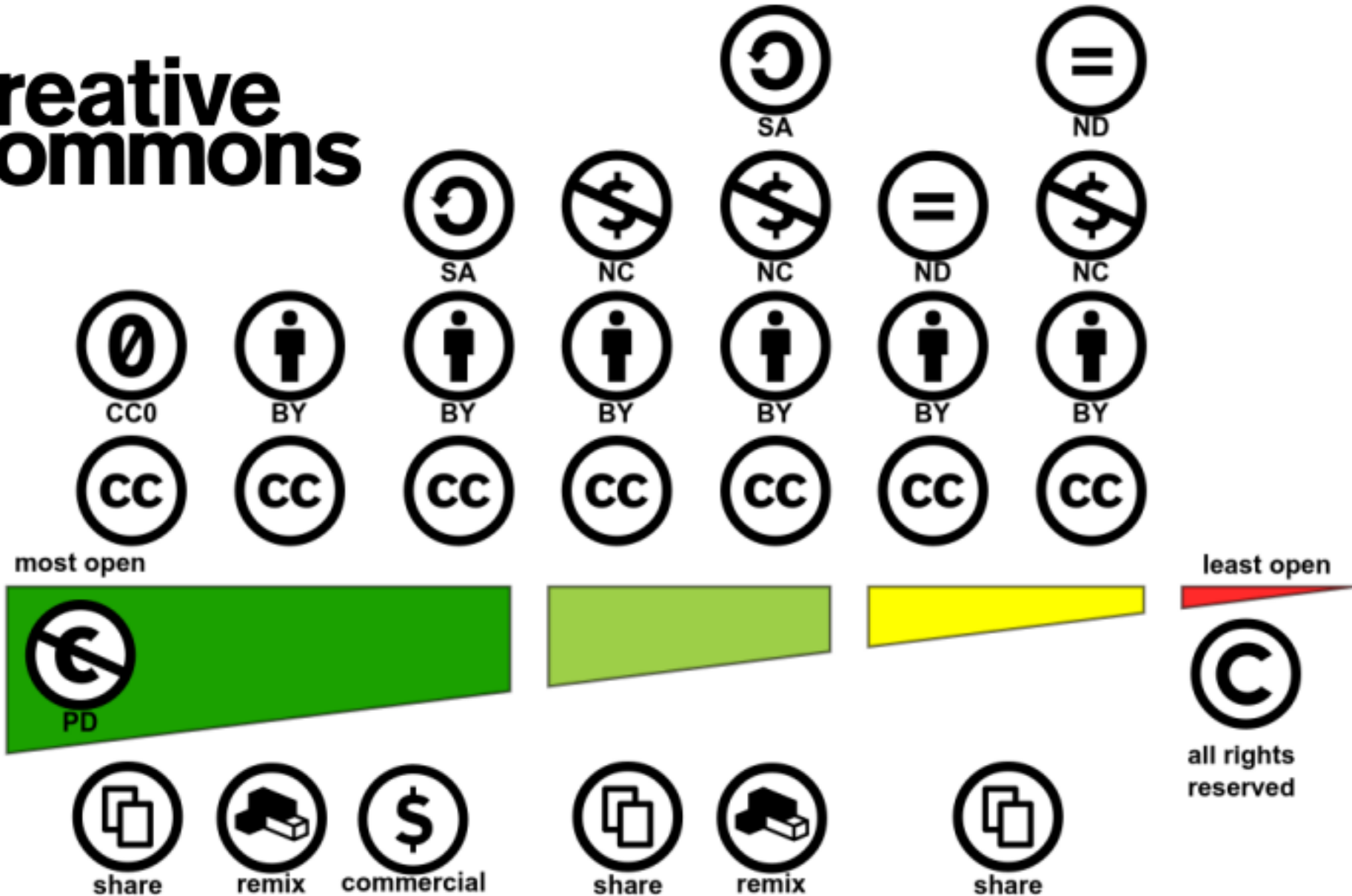
- Credit must be given to the creator
- Adaptations must be shared under the same terms



- Credit must be given to the creator



- is a public dedication tool, which allows creators to give up their copyright and put their works into the worldwide public domain.



Reusable code?

Licenses for code:

MIT (Expat)

Apache 2.0

GNU (L)GPL 3.0+

Use **Creative Commons Generator**:

<https://creativecommons.org/choose/?lang=en>

Choose an **open source license**:

<https://choosealicense.com/>

5 things to remember about archiving data:

1. If you put your data in an archive, you get a **DOI**
2. Not all archives offer **curation** – if the data have a long term value, choose an archive with curation
3. Add a **license** to your dataset and code so that others know what they can and cannot do with the data
4. If you want to share your **code**, archive a frozen, citeable version (e.g. via GitHub - Zenodo connection)
5. Without descriptions, documentation and structure, your data are most probably not reusable

Thank you!

Contact us at:

research-data@uio.no

Resources:

FAIR: <https://www.force11.org/fairprinciples>

DataverseNO: <https://dataverse.no>

NSD: <https://www.nsd.no/en/archiving-research-data/>

Zenodo: <https://zenodo.org>

Figshare: <https://figshare.com>

OSF <https://osf.io>

Re3data: <https://www.re3data.org/>

FAIRsharing: <https://fairsharing.org/>

Github citable code: <https://guides.github.com/activities/citable-code/>

Creative Commons Licenses: <https://creativecommons.org>

UiO Research data management <https://www.uio.no/english/for-employees/support/research/research-data-management/>