

4eu+

The FAIR principles in science, technology and engineering

How to write a good Data Management Plan for FAIR research data

Cécile Arènes Sorbonne University

Falco Hüser University of Copenhagen

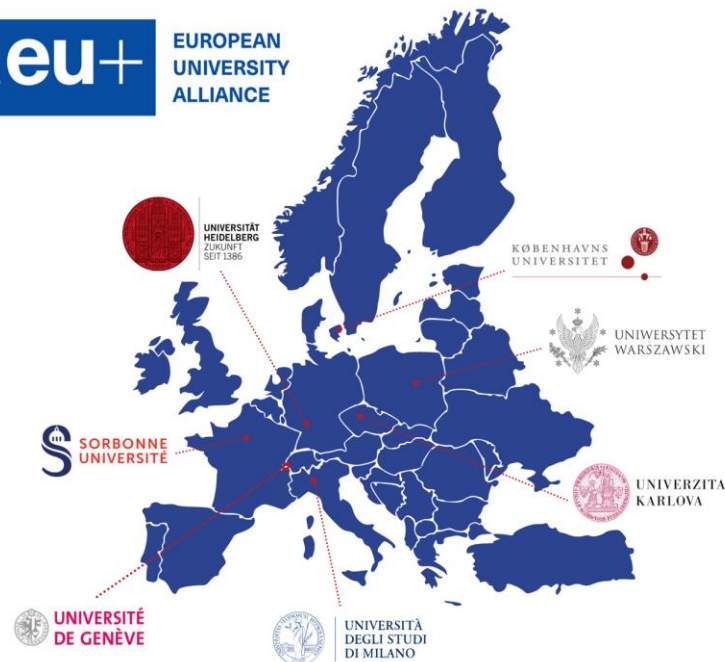
Asger Væring Larsen University of Copenhagen

Open for you! An introduction series to Open Science II 26 June 2023



4EU+ Alliance and Open Science

- 4EU+ is a **transnational strategic university association**.
- Aim: **Strengthen the European vision of deepened cooperation and mutual enrichment in research and teaching**
- Open Science is an integral part of this.
- **Two 4EU+ projects currently work on Open Science**





Open for YOU!
an introduction
series to Open Science

online training sessions from February to July 2023

Outline

Part 1: Introduction

- The FAIR principles
- Data Management Plans (DMP's)
- The Science Europe DMP template

Part 2: Examples

- Sharing Research Data
- Research Data Repositories
- Metadata Standards

Part 3: FAIR methods and tools

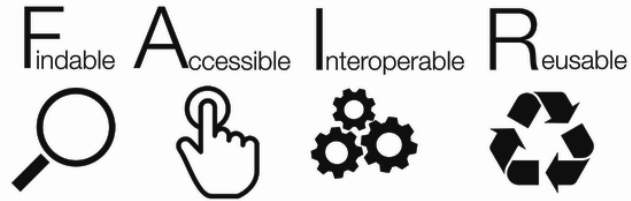
- FAIR assessment
- 5-star Open Data
- RDF – Resource Description Framework

Part 1: Introduction

- The FAIR principles
- Data Management Plans (DMP's)
- The Science Europe DMP template

Cécile Arènes Sorbonne University

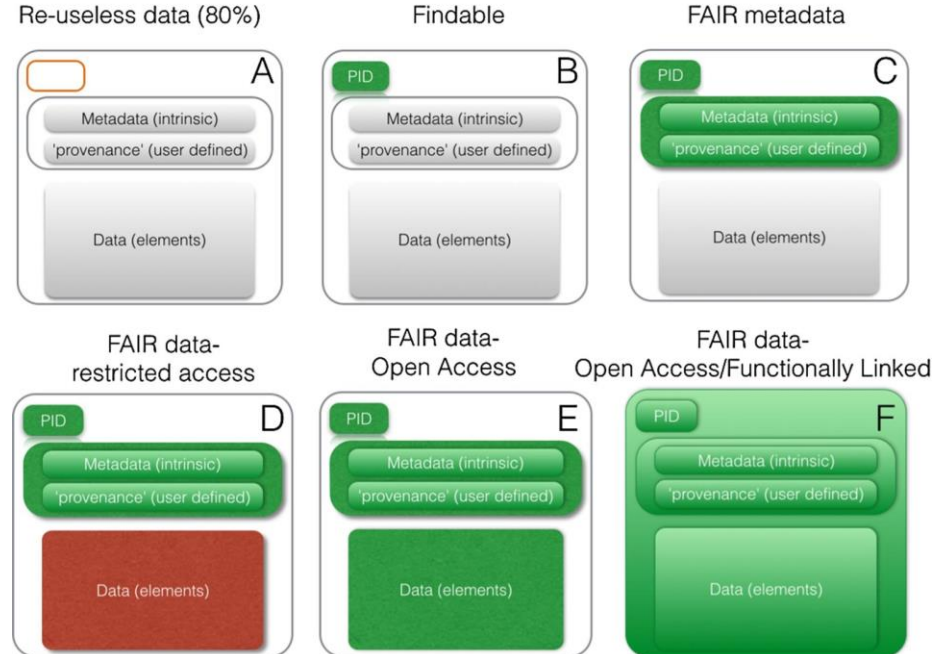
Cecile.Arenes@scd.upmc.fr




Aim of the FAIR principles:

- to offer several types of data sharing, with at least one persistent identifier and a standardized, sourced description (metadata)
- the dataset must be findable, while the data can be protected if necessary.

Data as increasingly FAIR Digital Objects



F_{indable} A_{ccessible} I_{nteroperable} R_{eusable}



principles in a nutshell

FAIR principles:

<https://force11.org/info/the-fair-data-principles/>



Image Credits:

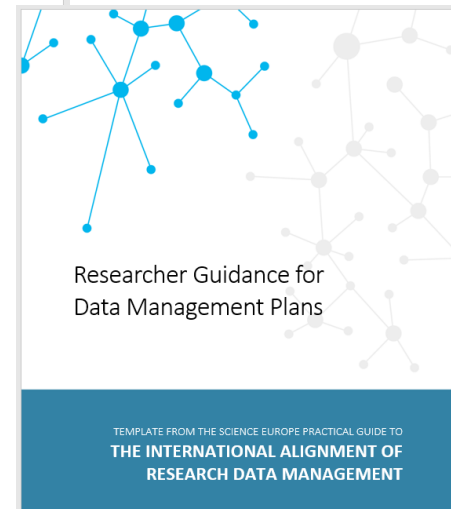
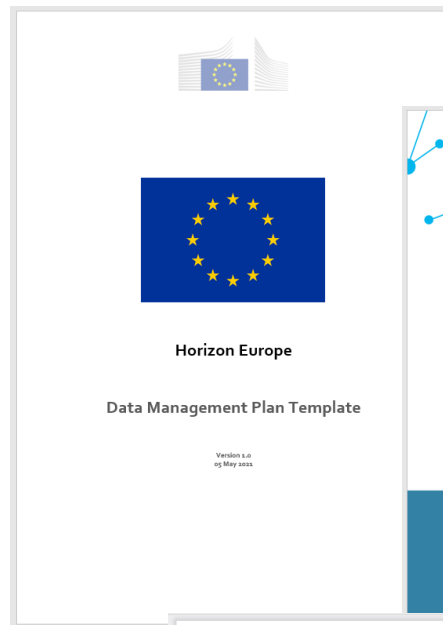
Logo [SangyaPundir](#), CC-BY-SA 4.0

Infographic [ANDS](#) CC-BY 4.0

A data management plan?

A data management plan (DMP) is a written document that **describes the data you expect to acquire or generate** during the course of a research project, how you will **manage, describe, analyze, and store those data**, and what mechanisms you will use at the end of your project to **share and preserve your data**.

<https://library.stanford.edu/research/data-management-services/data-management-plans>



Checklist for a Data Management Plan, v4.0

Please cite as: DCC. (2013). *Checklist for a Data Management Plan*. v.4.0. Edinburgh: Digital Curation Centre. Available online: <http://www.dcc.ac.uk/resources/data-management-plans>

DCC Checklist	DCC Guidance and questions to consider
Administrative Data	
ID	A pertinent ID as determined by the funder and/or institution.
Funder	State research funder if relevant
Grant Reference Number	Enter grant reference number if applicable [POST-AWARD DMPs ONLY]
Project Name	If applying for funding, state the name exactly as in the grant proposal.
Project Description	<p>Questions to consider:</p> <ul style="list-style-type: none"> - What is the nature of your research project? - What research questions are you addressing? - For what purpose are the data being collected or created? <p>Guidance:</p> <p>Briefly summarise the type of study (or studies) to help others understand the purposes for which the data are being collected or created.</p>

The DMP: timeline

- 1966: sketches of DMP in the aeronautical field
- 1973: NASA publishes a [technical report](#) that resembles a DMP.
- 2007: the Wellcome Trust (UK), now a member of Plan S, requires DMPs for the projects it funds
- 2007: [OECD guidelines](#)
- 2011: implementation of DMP by the National Science Foundation (USA) for funded projects.
- 2014: DMP for projects funded under H2020

Science Europe DMP template

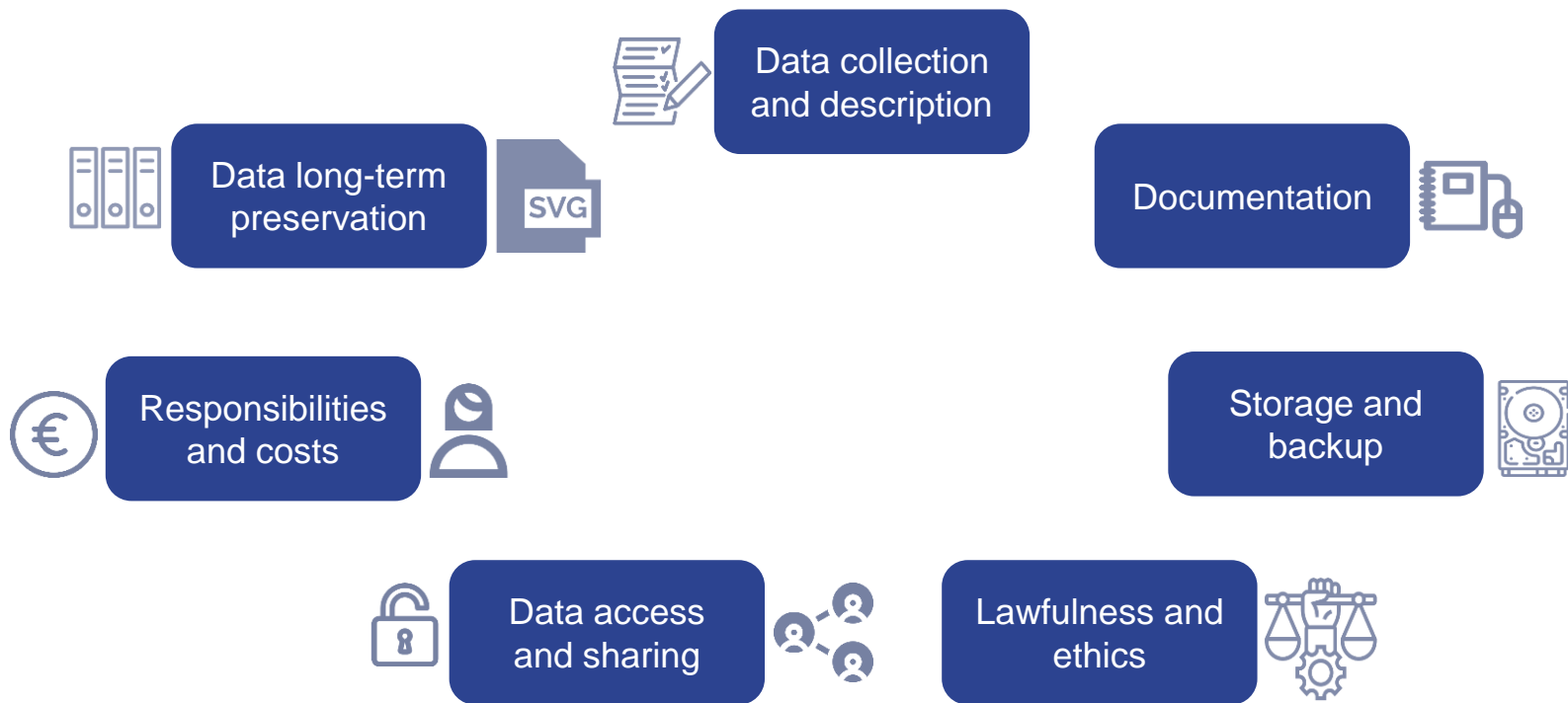
A user-friendly model that follows the project timeline



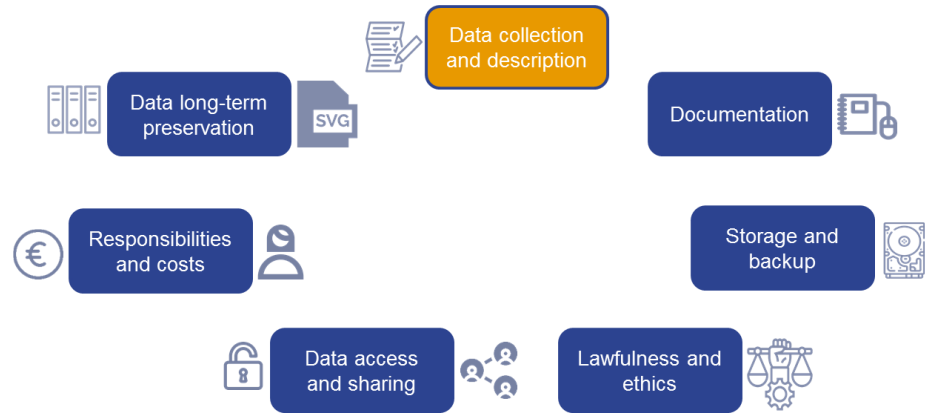
Researcher Guidance for
Data Management Plans

TEMPLATE FROM THE SCIENCE EUROPE PRACTICAL GUIDE
THE INTERNATIONAL ALIGNMENT C

Science Europe DMP in brief



Main fields of the Science Europe template – 1



General information

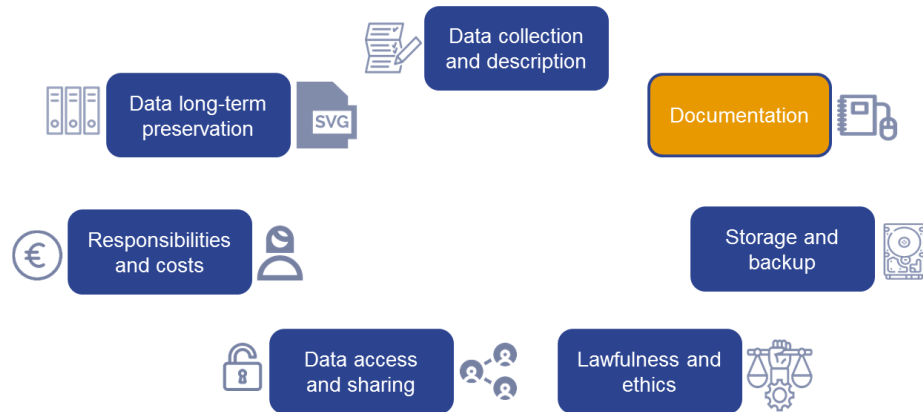
Administrative information

1. DATA DESCRIPTION AND COLLECTION OR RE-USE OF EXISTING DATA

1a. How will new data be collected or produced and/or how will existing data be re-used?

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

Main fields of the Science Europe template – 2

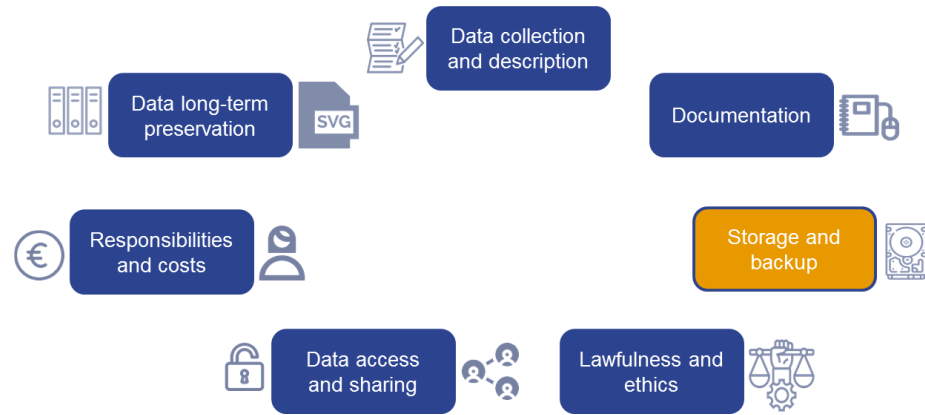


2. DOCUMENTATION AND DATA QUALITY

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

2b. What data quality control measures will be used?

Main fields of the Science Europe template – 3

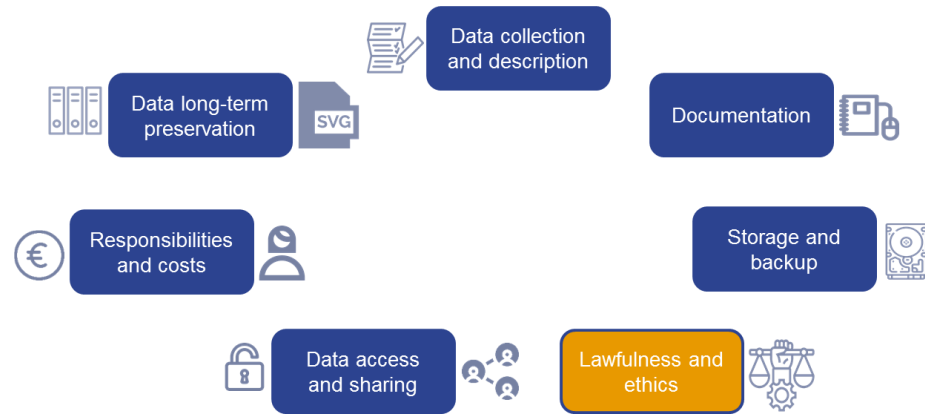


3. STORAGE AND BACKUP DURING THE RESEARCH PROCESS

3a. How will data and metadata be stored and backed up during the research?

3b. How will data security and protection of sensitive data be taken care during the research?

Main fields of the Science Europe template – 4



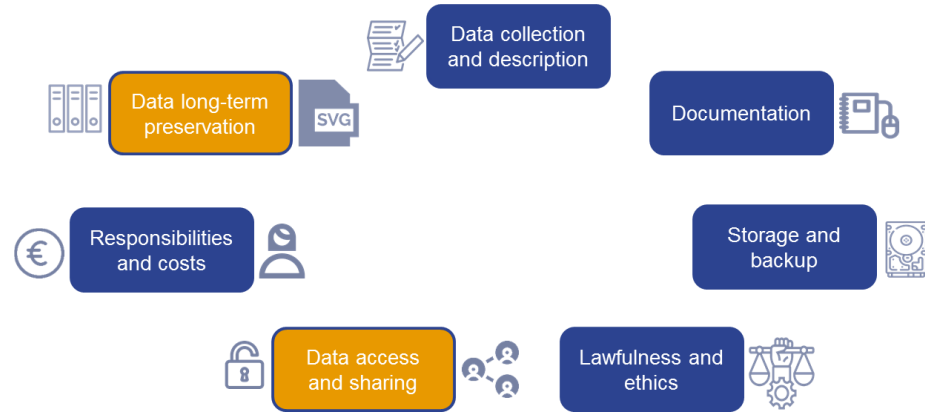
4. LEGAL AND ETHICAL REQUIREMENTS, CODE OF CONDUCT

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

Main fields of the Science Europe template – 5



5. DATA SHARING AND LONG-TERM PRESERVATION

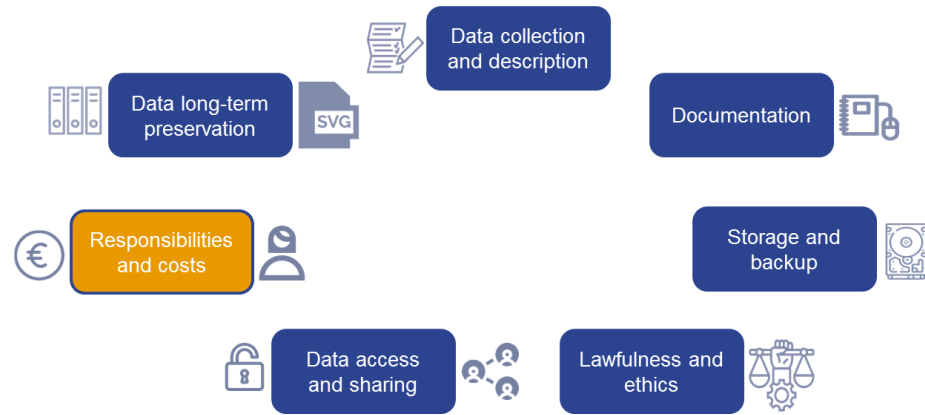
5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

5c. What methods or software tools are needed to access and use data?

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

Main fields of the Science Europe template – 6



6. DATA MANAGEMENT RESPONSIBILITIES AND RESOURCES

6a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

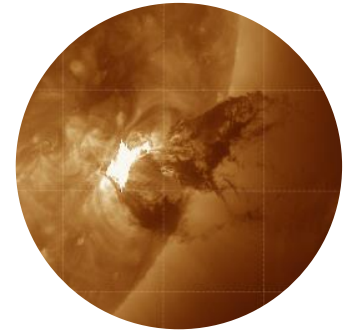
6b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

Part 2: Examples

- Sharing Research Data
- Research Data Repositories
- Metadata Standards

Falco Hüser University of Copenhagen

falh@kb.dk



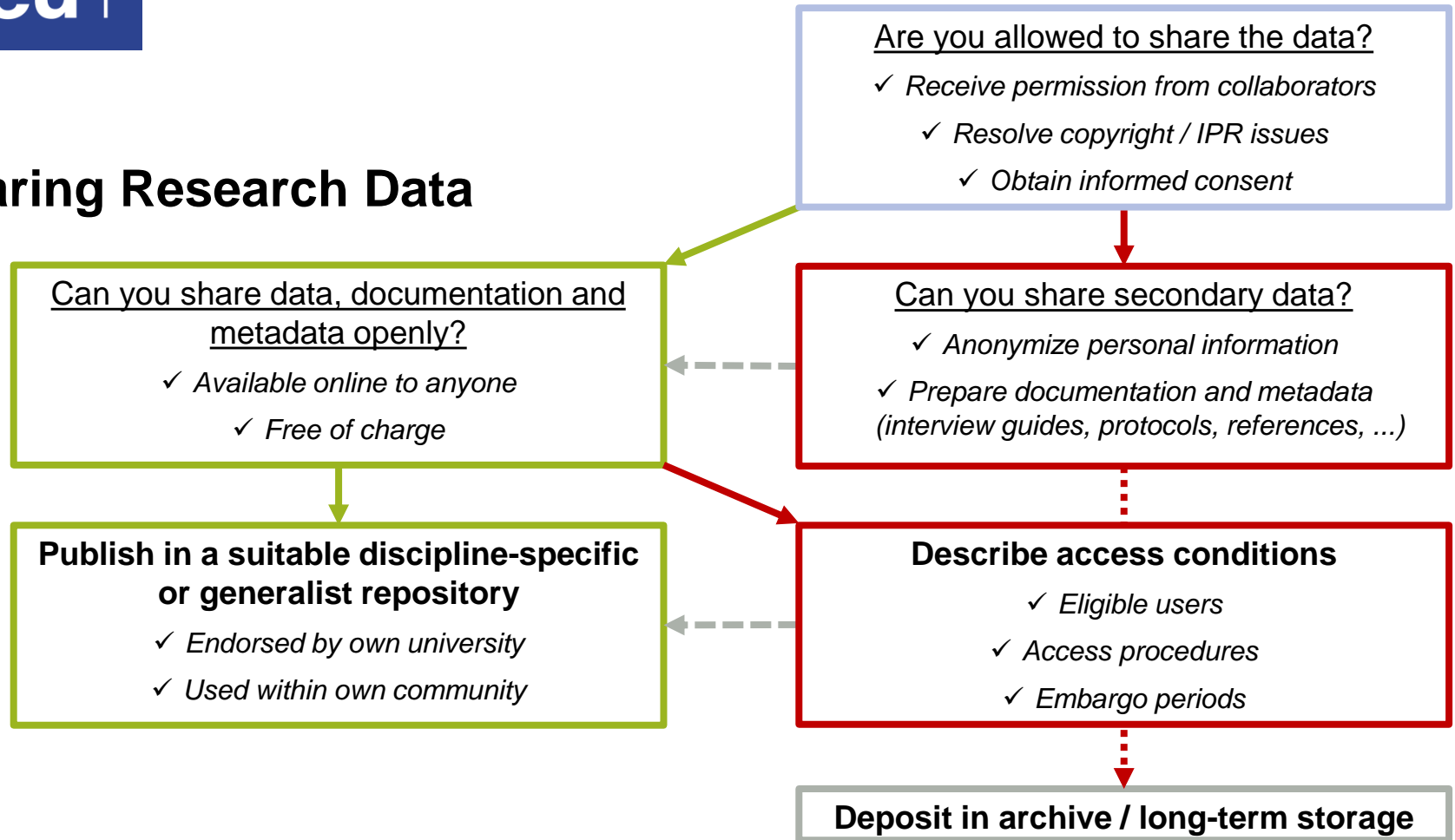
Sharing Research Data

5a How and when will data be shared?

Are there possible restrictions to data sharing or embargo reasons?

- Explain how the data will be discoverable and shared (for example by deposit in a trustworthy data repository, indexed in a catalogue, use of a secure data service, direct handling of data requests, or use of another mechanism).
- Outline the plan for data preservation and give information on how long the data will be retained.
- Explain when the data will be made available. Indicate the expected timely release. Explain whether exclusive use of the data will be claimed and if so, why and for how long. Indicate whether data sharing will be postponed or restricted for example to publish, protect intellectual property, or seek patents.
- Indicate who will be able to use the data. If it is necessary to restrict access to certain communities or to apply a data sharing agreement, explain how and why. Explain what action will be taken to overcome or to minimise restrictions.

Sharing Research Data



DMP Example

“Fully anonymizable data will be made openly available. This includes raw and processed EEG and fNIRS recordings, eyetracking data, and excel files containing coding of children's behavior and looking-durations.”

“Video recordings of participants contain identifiable information, and will not be made accessible for data protection reasons.”

“Non-anonymizable data will be kept on file by the researcher on a secure drive provided by the host institution for at least 10 years after conclusion of the data and will only be shared with researchers directly involved in the project (in accordance with data protection regulations).”

Research Data Repositories

Heidelberg University: <https://heidata.uni-heidelberg.de/>

Sorbonne University: <https://recherche.data.gouv.fr/>

University of Milan: <https://dataverse.unimi.it/>

University of Warsaw: <https://repod.icm.edu.pl/>

University of Copenhagen: <https://erda.ku.dk/>

University of Geneva: <https://yareta.unige.ch/>

Charles University: *no institutional / national repository*

heiDATA | Heidelberg
Open Research Data



recherche.data.gouv.fr



UNIVERSITÀ
DEGLI STUDI
DI MILANO

The
Dataverse[®]
Project



RepOD

Repository for Open Data



Electronic Research Data Archive
University of Copenhagen



YARETA

Research Data Repositories

Registry of Research Data Repositories: <https://www.re3data.org/>



Repository offers open access



Repository offers restricted access



Repository offers closed access



Repository issues DOI's



Repository provides reuse licenses



Repository is certified



DMP Example

Mass spectrometry proteomics data will be deposited to the PRoteomics IDentification Database PRIDE (<https://www.ebi.ac.uk/pride/>).

- PRIDE facilitates free and unhindered access to all datasets after publication.
- PRIDE is part of the ELIXIR infrastructure and regarded as well-established standard repository in the field.
- All datasets deposited in PRIDE are made available under Creative Commons Public Domain (CC0).
- All datasets deposited in PRIDE receive unique dataset identifiers (PXD#####).
- All data submitted to PRIDE are being reviewed by expert bio-curators.



DMP Example



Samples will be registered with the SESAR "System for Earth Sample Registration" (<http://www.geosamples.org/>).

“SESAR is a community platform that helps make samples more discoverable, accessible, and reusable, and connects samples with the knowledge ecosystem derived from them.”

“Every sample submitted to the SESAR index is assigned an IGSN, which gives the sample a globally unique and persistent identifier.”

By default, “sample metadata are publically available immediately upon registration.”

Metadata and Documentation

2a What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

- Indicate which metadata will be provided to help others identify and discover the data.
- Indicate which metadata standards (for example DDI, TEI, EML, MARC, CMDI) will be used.
- Use community metadata standards where these are in place.
- Indicate how the data will be organised during the project, mentioning for example conventions, version control, and folder structures.
- Consistent, well-ordered research data will be easier to find, understand, and re-use.
- Consider what other documentation is needed to enable re-use. This may include information on the methodology used to collect the data, analytical and procedural information, definitions of variables, units of measurement, and so on.
- Consider how this information will be captured and where it will be recorded for example in a database with links to each item, a 'readme' text file, file headers, code books, or lab notebooks.

Metadata



<https://opengeospatial.github.io/e-learning/metadata/>



<https://www.etsy.com/dk-en/shop/DonBurns27>

Nutrition Facts

Serving Size 172 g

Amount Per Serving

Calories 200 Calories from Fat 8

% Daily Value*

Total Fat 1g 1%

Saturated Fat 0g 1%

Trans Fat

Cholesterol 0mg 0%

Sodium 7mg 0%

Total Carbohydrate 36g 12%

Dietary Fiber 11g 45%

Sugars 6g

Protein 13g

Vitamin A 1% • Vitamin C 1%

Calcium 4% • Iron 24%

*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs.

!nutritionData.com

TOPICS

- Chemical compounds and components
- Chemical elements
- Carbon based materials
- Graphene
- Heterocyclic compounds
- Thermoelectric effects
- Electronic transport

Original language English

Article number 214302

Journal The Journal of Chemical Physics

Volume 143

Issue number 21

Number of pages

ISSN

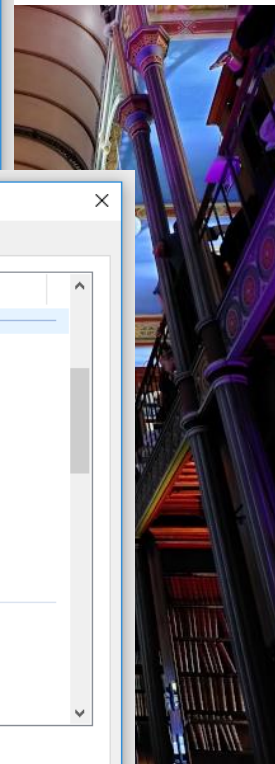
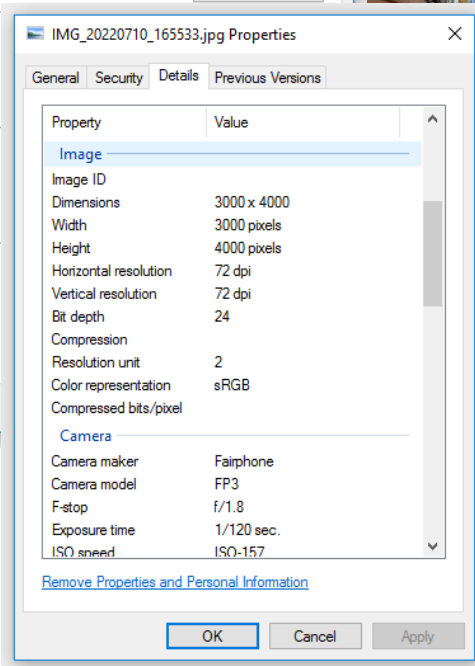
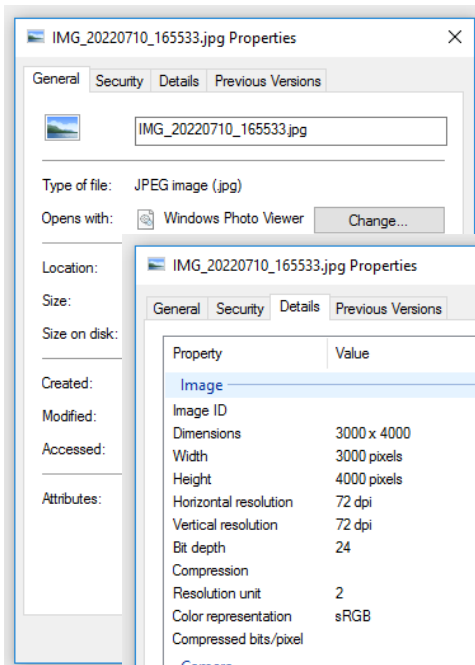
DOIs

Publication status

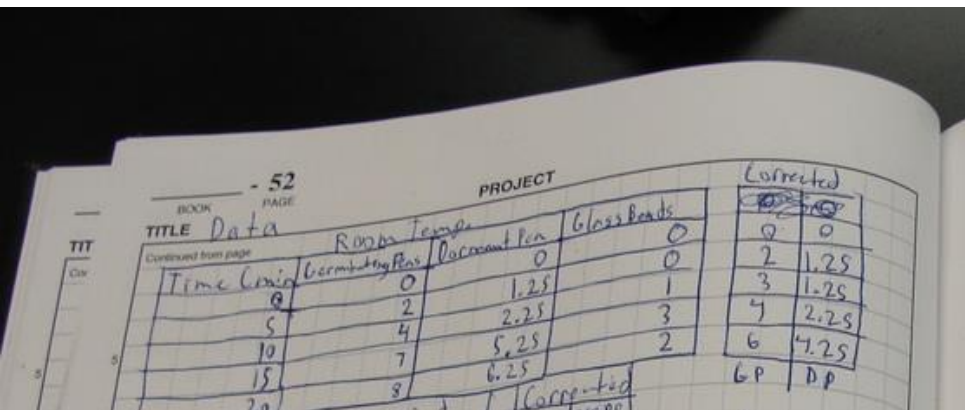
REFERENCES

1. H. Song, M. A. Reed, and T. Lee, "Single molecule electronic devices," Adv. Mater. **23**, 1583–1608 (2011).

<https://doi.org/10.1002/adma.201004291>, [Google Scholar](#), [Crossref](#)



533.jpg



DMP Example

Each dataset in a data repository will be described with metadata using a README file (.txt or .pdf). The README file consists of three parts:

Project-level descriptions:

- explain the aims and research questions of the study, the hypothesis, the measurement equipment and experimental setup, the used methodology and type of data;

File-level documentation:

- explain how all the files that make up a dataset relate to one another;

Item-level documentation explaining the names of the variables and the meanings of those variables.

- explain which data-files contain which variables and what these variables represent;

Replication Data for: Spawning time in adult polar cod (Boreogadus saida) altered by crude oil exposure, independent of food availability

Version 2.0



Strople, Leah C.; Vieweg, Ireen; Nahrgang, Jasmine; Yadetie, Fekadu; Odei, Derrick Kwame; Thorsen, Anders; Karlsen, Odd André; Goksøyr, Anders; Sørensen, Lisbet; Sarno, Antonio; Hansen, Bjørn Henrik; Frantzen, Marianne; Hansen, Øyvind; Puvanendran, Velmurugu, 2023. "Replication Data for: Spawning time in adult polar cod (Boreogadus saida) altered by crude oil exposure, independent of food availability", <https://doi.org/10.18710/59XOI4>, DataverseNO, V2

[Cite Dataset](#)
[Learn about Data Citation Standards.](#)
[Access Dataset](#)
[Contact Owner](#)
[Share](#)
[Dataset Metrics](#)

1 Download

1 to 8 of 8 Files


[00_README_AeN_FRAM.txt](#)

Plain Text - 27.5 KB

Published Jan 9, 2023

1 Download

MD5: f28...b6b

Description of data and methods


[AeN_FRAM_DEGs_Differentially_Expressed_Genes.xlsx](#)

MS Excel Spreadsheet - 150.0 KB

Published May 10, 2023

0 Downloads

MD5: f2b...42e

List of differentially expressed genes from female polar cod taken on day 47 of the experiment.
Comparison completed between five-high-feed oil-exposed females and five high-feed control females.


[AeN_FRAM_dryweight_measurements.csv](#)

Comma Separated Values - 6.3 KB

Published Jan 9, 2023

0 Downloads

MD5: e34...c03

Samples of fish that were taken to determine the wet somatic weight to dry somatic weight ratio a subset of gonads were also dried to determine the wet to dry weight ratio


[AeN_FRAM_egg_PAH.csv](#)

Comma Separated Values - 37.8 KB

Published Jan 9, 2023

0 Downloads

MD5: adb...547

This file contains PAH concentrations in ng/g detected in egg samples. Samples were taken 47 days into the experimental period


[AeN_FRAM_experimental_photoperiod.csv](#)

Comma Separated Values - 879 B

Published Jan 9, 2023

0 Downloads

MD5: fbc...87a

Photoperiod during the experimental period



DMP Example

EMO BON data are accompanied by rich and rigorous metadata that include, but are not limited to, information on the where, when, and how the samples were collected (**Observatory Metadata** and **Sampling Metadata**). Additional *Complementary (Meta)data* include the environmental variables measured during a sampling event and the methodologies used to collect measure them. Information on the laboratory analyses of the data, such as the DNA extraction method, the yields and the library preparation are collected as **Analysis Metadata**. The quality controlled bioinformatics procedures following sequencing to produce the Quality-controlled Sequence Data are documented as **Post-sequencing Metadata**. The Source Material Identifier is included in the metadata records and links together all the information collected as metadata.

Metadata Standards

- Provide a common 'language' for the community.
- Enable interoperability across disciplines (and sectors).
- Are ideally described in a citable online resource.
- Should be readable by humans and machines.
- Can be embedded in file formats.

Vocabularies provide unambiguous definitions for individual metadata elements.

Taxonomies structure metadata elements in a hierarchy.

Ontologies contain relations between metadata elements.

Index of subjects

Multidisciplinary

Science

Atmospheric sciences

Climatology

Meteorology

Biological sciences

Biochemistry

Biochemicals

Proteins

Metabolism

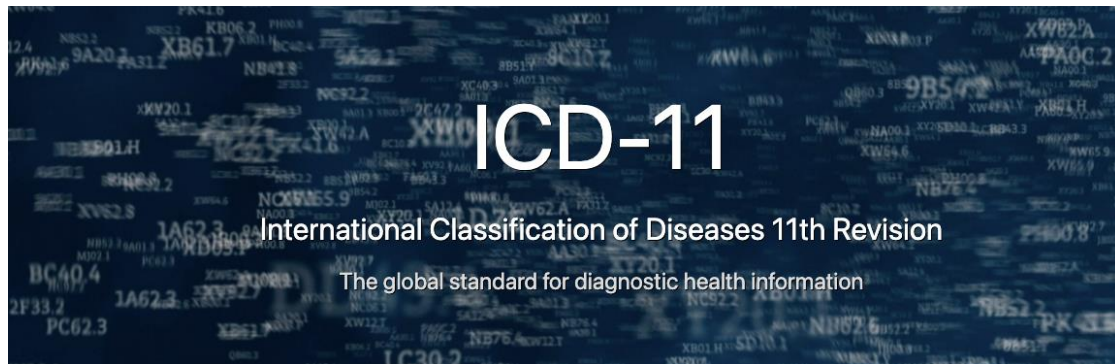
Biology

Biophysics

Cell biology

Genome

Example



5A10 Type 1 diabetes mellitus

All ancestors up to top

- 05 Endocrine, nutritional or metabolic diseases
 - Endocrine diseases
 - Diabetes mellitus
 - 5A10 Type 1 diabetes mellitus

Hide ancestors (X)

Description

Diabetes mellitus type 1 (type 1 diabetes, T1DM, formerly insulin dependent or juvenile diabetes) is a form of diabetes mellitus that results from destruction of insulin-producing beta cells, mostly by autoimmune mechanisms. The subsequent lack of insulin leads to increased blood and urine glucose.

Exclusions

- Type 2 diabetes mellitus (5A11)
- Diabetes mellitus, other specified type (5A13)
- Diabetes mellitus in pregnancy (JA63)

Coded Elsewhere

- Pre-existing type 1 diabetes mellitus in pregnancy (JA63.0)



**World Health
Organization**

<https://icd.who.int/en>



Example



Drosophila melanogaster

Taxonomy ID: 7227 (for references in articles please use NCBI:txid7227)

current name

Drosophila melanogaster Meigen, 1830

homotypic synonym: *Sophophora melanogaster* (Meigen, 1830)

includes: *Diptera* sp. DNAS-2A9-224646

Genbank common name: **fruit fly**

NCBI BLAST name: **flies**

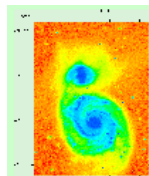
Rank: **species**

Genetic code: [Translation table 1 \(Standard\)](#)

Mitochondrial genetic code: [Translation table 5 \(Invertebrate Mitochondrial\)](#)

[Lineage \(full\)](#)

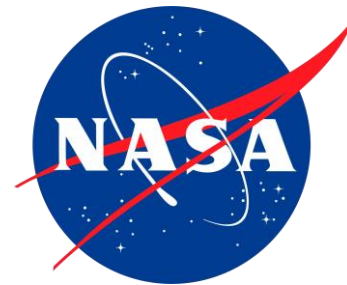
[cellular organisms](#); [Eukaryota](#); [Opisthokonta](#); [Metazoa](#); [Eumetazoa](#); [Bilateria](#); [Protostomia](#); [Ecdysozoa](#); [Panarthropoda](#); [Arthropoda](#); [Mandibulata](#); [Pancrustacea](#); [Hexapoda](#); [Insecta](#); [Dicondylia](#); [Pterygota](#); [Neoptera](#); [Endopterygota](#); [Diptera](#); [Brachycera](#); [Muscomorpha](#); [Eremoneura](#); [Cyclorrhapha](#); [Schizophora](#); [Acalypratae](#); [Ephydroidea](#); [Drosophilidae](#); [Drosophilinae](#); [Drosophilini](#); [Drosophila](#); [Sophophora](#); [melanogaster group](#); [melanogaster subgroup](#)



FITS

Flexible Image Transport System

The Astronomical
Image and Table Format

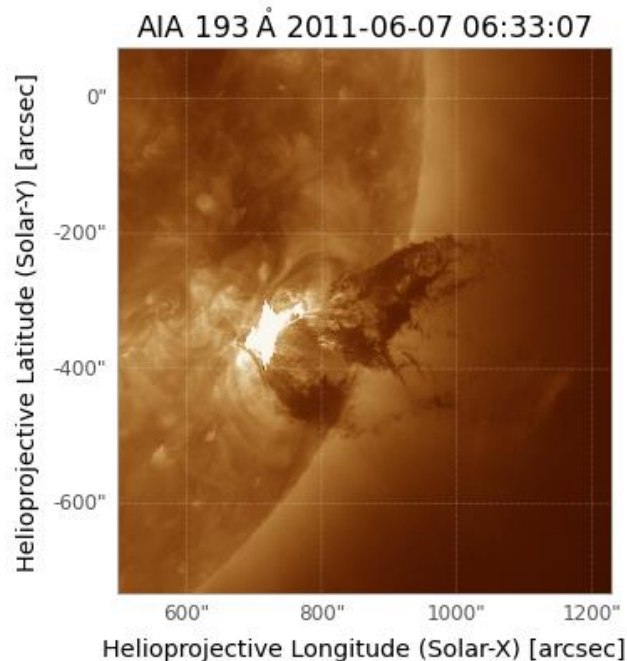


Example

Table 27: Spectral reference systems.

Value	Definition
'TOPOCENT'	Topocentric
'GEOCENTR'	Geocentric
'BARYCENT'	Barycentric
'HELIOCEN'	Heliocentric
'LSRK'	Local standard of rest (kinematic)
'LSRD'	Local standard of rest (dynamic)
'GALACTOC'	Galactocentric
'LOCALGRP'	Local Group
'CMBDIPOL'	Cosmic-microwave-background dipole
'SOURCE'	Source rest frame

https://fits.gsfc.nasa.gov/standard40/fits_standard40aa-le.pdf



Part 3: FAIR methods and tools

- FAIR assessment
- 5-star Open Data
- RDF – Resource Description Framework

Asger Væring Larsen University of Copenhagen

avla@kb.dk

April 16, 2023

Dataset Open Access

A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration

Bandana, Juan M.; Tekumalla, Ramya; Wang, Guanyu; Yu, Jingyuan; Liu, Tu; Ding, Yuning; Artemova, Katya; Tutubalina, Elena; Chowell, Gerardo

Version 162 of the dataset. NOTES: Data for 3/15 - 3/18 was not extracted due to unexpected and unannounced downtime of our university infrastructure. We will try to backfill those days by next release. FUTURE CHANGES: Due to the imminent paywalling of Twitter's API access this might be the last full update of this dataset. If the API access is not blocked, we will be stopping updates for this dataset with release 165 - a bit more than 3 years after our initial release. It's been a joy seeing all the work that uses this resource and we are glad that so many found it useful.

The dataset files: `full_dataset.tsv.gz` and `full_dataset_clean.tsv.gz` have been split in 1 GB parts using the Linux utility called Split. So make sure to join the parts before unzipping. We had to make this change as we had huge issues uploading files larger than 2GB's (hence the delay in the dataset releases). The peer-reviewed publication for this dataset has now been published in *Epidemiologia an MDPI journal*, and can be accessed here: <https://doi.org/10.3390/epidemiologia2030024>. Please cite this when using the dataset.

Due to the relevance of the COVID-19 global pandemic, we are releasing our dataset of tweets acquired from the Twitter Stream related to COVID-19 chatter. Since our first release we have received additional data from our new collaborators, allowing this resource to grow to its current size. Dedicated data gathering started from March 11th yielding over 4 million tweets a day. We have added additional data provided by our new collaborators from January 27th to March 27th, to provide extra longitudinal coverage. Version 10 added ~1.5 million tweets in the Russian language collected between January 1st and May 8th, gracefully provided to us by: Katya Artemova (NRU HSE) and Elena Tutubalina (KFU). From version 12 we have included daily hashtags, mentions and emojis and their frequencies the respective zip files. From version 14 we have included the tweet identifiers and their respective language for the clean version of the dataset. Since version 20 we have included language and place location for all tweets.

The data collected from the stream captures all languages, but the higher prevalence are: English, Spanish, and French. We release all tweets and retweets on the `full_dataset.tsv` file (1,395,222,801 unique tweets), and a cleaned version with no retweets on the `full_dataset-clean.tsv` file (361,748,721 unique tweets). There are several practical reasons for us to leave the retweets, tracing important tweets and their dissemination is one of them. For NLP tasks we provide the top 1000 frequent terms in `frequent_terms.csv`, the top 1000 bigrams in `frequent_bigrams.csv`, and the top 1000 trigrams in `frequent_trigrams.csv`. Some general statistics per day are included for both datasets in the `full_dataset-statistics.tsv` and `full_dataset-clean-statistics.tsv` files. For more statistics and some visualizations visit: <http://www.panacealab.org/covid19/>

More details can be found (and will be updated faster at: https://github.com/thepanacealab/covid19_twitter) and our

252,328

views

209,342

downloads

[See more details...](#)

Indexed in

OpenAIRE

Publication date:

April 16, 2023

DOI:

DOI: 10.5281/zenodo.7834392

Keyword(s):

social media twitter nlp covid-19 covid19

Published in:

Epidemiologia: 2 pp. 315-324 (3).

Related identifiers:Continued by
<http://www.panacealab.org/covid19/> (Other)Supplement to
<https://arxiv.org/abs/2004.03688> (Preprint)**Alternate identifiers:**10.3390/epidemiologia2030024 (Journal article)
https://github.com/thepanacealab/covid19_twitter
(Software)**Communities:**BioHackathon
Coronavirus Disease Research Community -
COVID-19

Metadata

DOI



As always, the tweets distributed here are only tweet identifiers (with date and time added) due to the terms and conditions of Twitter to re-distribute Twitter data ONLY for research purposes. They need to be hydrated to be used.

This dataset will be updated bi-weekly at least with additional tweets, look at the github repo for these updates.
Release: We have standardized the name of the resource to match our pre-print manuscript and to not have to update it every week.

License (for files):
[Other \(Public Domain\)](#)

License

Files

Preview

emojis.zip

The previewer is not showing all the files

- extracted_elements
 - emojis
 - 2020-01-04_clean-emoji_char.tsv 11 Bytes
 - 2020-01-04_clean-emoji_text.tsv 21 Bytes
 - 2020-01-06_clean-emoji_char.tsv 1 Byte
 - 2020-01-06_clean-emoji_text.tsv 1 Byte
 - 2020-01-08_clean-emoji_char.tsv 1 Byte
 - 2020-01-08_clean-emoji_text.tsv 1 Byte
 - 2020-01-09_clean-emoji_char.tsv 1 Byte
 - 2020-01-09_clean-emoji_text.tsv 1 Byte
 - 2020-01-10_clean-emoji_char.tsv 1 Byte
 - 2020-01-10_clean-emoji_text.tsv 1 Byte
 - 2020-01-11_clean-emoji_char.tsv 29 Bytes
 - 2020-01-11_clean-emoji_text.tsv 104 Bytes

Files (16.1 GB)

Name	Size	Preview	Download
emojis.zip	15.1 MB	Preview	Download
md5:794aa07e49f5edf3ed72d552321bb2f5			
frequent_bigrams.csv	17.9 kB	Preview	Download
md5:c7019423b59057512d7c65777efa2067			
frequent_terms.csv	11.2 kB	Preview	Download
md5:bfa25849251420d474671f6ba8dae969			
frequent_trigrams.csv	25.0 kB	Preview	Download
md5:b1f1...d4f0000f0>50c501016c487d1			

Versions

Version 162	Apr 16, 2023
10.5281/zenodo.7834392	
Version 161	Apr 9, 2023
10.5281/zenodo.7812326	
Version 160	Apr 2, 2023
10.5281/zenodo.7793560	
Version 159	Mar 26, 2023
10.5281/zenodo.7772372	
Version 158	Mar 19, 2023
10.5281/zenodo.7753101	

[View all 163 versions](#)

Cite all versions? You can cite all versions by using the DOI [10.5281/zenodo.3723939](https://doi.org/10.5281/zenodo.3723939). This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)

Share



Cite as

Banda, Juan M., Tekumalla, Ramya, Wang, Guanyu, Yu, Jingyuan, Liu, Tuo, Ding, Yuning, Artemova, Katya, Tutubalina, Elena, & Chowell, Gerardo. (2023). A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration [Data set]. In *Epidemiologia* (Version 162, Vol. 2, Number 3, pp. 315–324). Zenodo. <https://doi.org/10.5281/zenodo.7834392>

Start typing a citation style...



F-UJI

Automated FAIR Data
Assessment Tool

F-UJI is a web service to programmatically assess FAIRness of research data objects at the dataset level based on the FAIRsFAIR Data Object Assessment Metrics [↪](#)

[Click here to assess a dataset](#)

<https://www.f-uji.net/>

FAIR assessment

F-UJI is a web service to programmatically assess FAIRness of research data objects (aka data sets) based on metrics developed by the [FAIRsFAIR](#) project.

Please use the form below to enter an identifier (e.g. DOI, URL) of the data set you wish to assess. Optionally you also can enter a metadata service (OAI-PMH, SPARQL, CSW) endpoint URI which F-UJI can use to identify additional information.

Research Data Object (URL/PID):*

[Settings](#)

▶ Start FAIR Assessment

[About](#)[Feedback](#)[Privacy Policy](#)[Terms of Use](#)[Legal Notice](#)

FAIR assessment

Disclaimer:

The test results shown here are based on preliminary data and code which still is under development. F-UI is rapidly evolving and not yet available in a productive environment.

[Click here to assess another data set](#)

Assessment Results:

Evaluated Resource:

A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration

✓ Save

↓ {JSON}

📄 New

FAIR level: ⓘ

advanced

Resource PID/URL:

<https://doi.org/10.5281/zenodo.7834392>

DataCite support:

enabled

Metric Version:

metrics_v0.5

Metric Specification:

<https://doi.org/10.5281/zenodo.4081213>

Software version:

2.2.5

Download assessment results:

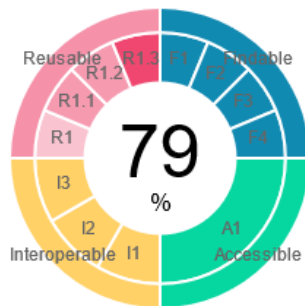
[{JSON}](#)

Save and share assessment results:

Saved assessments:

- FAIR 79% [2023-05-17 \(2.2.5\)](#) ⚙️

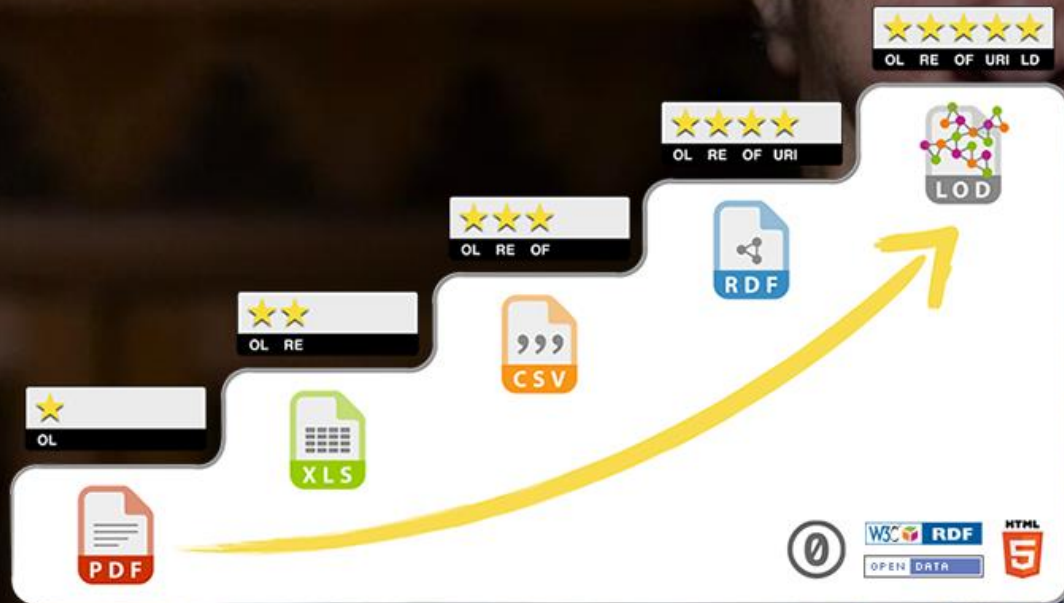
Summary:



	Score earned:		Fair level:
Findable:	7 of 7	🔄	advanced
Accessible:	3 of 3	🔄	advanced
Interoperable:	4 of 4	🔄	advanced
Reusable:	5 of 10	🔄	moderate

5 ★ OPEN DATA

Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a 5-star deployment scheme for Open Data. Here, we give examples for each step of the stars and explain costs and benefits that come along with it.



Below, we provide examples for each level of Tim's 5-star Open Data plan. The example data used throughout is *'the temperature forecast for Galway, Ireland for the next 3 days'*:

- | | | |
|-------|--|-------------|
| ★ | make your stuff available on the Web (whatever format) under an open license ¹ | example ... |
| ★★ | make it available as structured data (e.g., Excel instead of image scan of a table) ² | example ... |
| ★★★ | make it available in a non-proprietary open format (e.g., CSV instead of Excel) ³ | example ... |
| ★★★★ | use URIs to denote things, so that people can point at your stuff ⁴ | example ... |
| ★★★★★ | link your data to other data to provide context ⁵ | example ... |

RDF – Resource Description

Framework

Triples

Semantic data

Linked open data

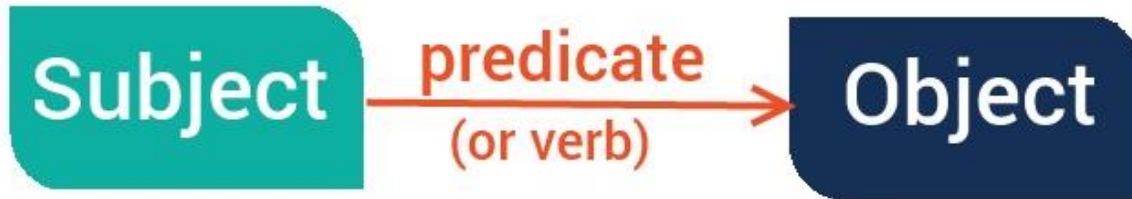
FAIR data points

Linked Data Platform

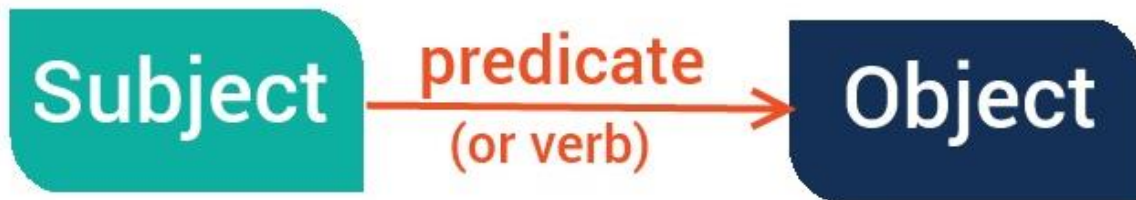
What is an RDF Triplestore?

The RDF triplestore is a type of graph database that stores data as a network of objects and uses inference to uncover new information out of existing relations. Its flexible and dynamic nature allows linking diverse data, indexing it for semantic search and enriching it via text analysis to build big knowledge graphs.





Subject	Predicate	Object
Wilma	hasSpouse	Fred
Fred	hasAge	25
Fred	livesIn	Bedrock



Subject	Predicate (propertyURL)	Object (valueURL)
Tokyo (http://example/ressource/tokyo)	hasArea (http://example/property/area)	2188 km2 (Literal)
	isInCountry (http://example/ressource/country)	Japan (http://example/ressource/Japan)

Making unFAIR data FAIR

Creating a file which contains the data AND ontology-controlled metadata as one package – a database

Upload to a triplestore

OpenRefine RDF extension

Examples of uses of RDF/Knowledge Graphs



Multi purpose:

[Google Knowledge Graph](#)

[Amazon's product graph](#)

[Dbpedia](#) (Open)

[Wikidata](#) (Open)

[Geonames](#) (Open)

[Yago](#) (Open)

"Real" research projects

The Human Genome project -> [Ensembl](#)

[The Linked Open Drug Data](#)

[BIO2RDF](#)

[Antimicrobial Compounds Database](#)

[Neurodata](#)



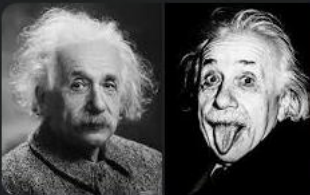
Albert Einstein

Teoretisk fysiker

Oversigt

Bøger

Videor



wikipedia.org

[https://da.wikipedia.org/wiki/Albert Einstein](https://da.wikipedia.org/wiki/Albert_Einstein)

Albert Einstein - Wikipedia

Albert Einstein (født 14. marts 1879) med en omfattende og banebrydende

Født: 14. marts 1879; Ulm, Württemberg
Nobelpris: Fysik 1921

Baggrund · Patentkontor · Einstein b

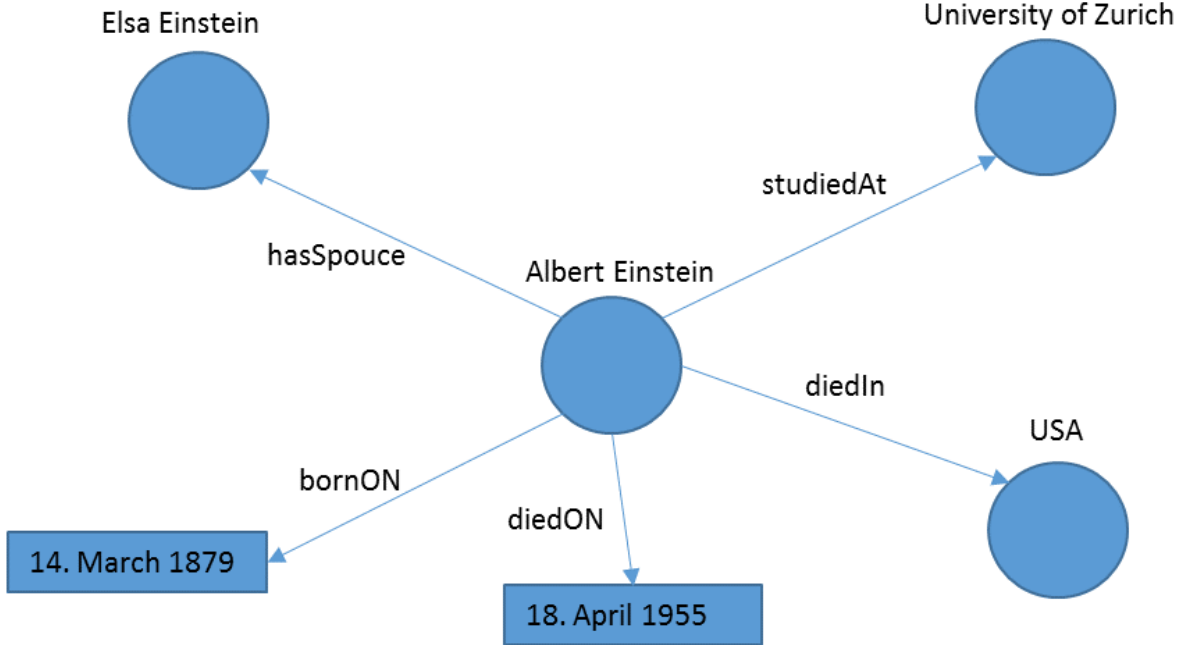
Folk spørger også om :

Hvor mange IQ har Albert Einstein?

Hvad er Albert Einstein kendt for?

Hvad har Einstein opfundet?

Hvor mange børn fik Einstein?



Feedback

Se mere →

Examples of FAIR/graph tools



Tools:

[OpenRefine](#) can model RDF data

[Neo4j](#) creates graphs

[Cedar Workbench](#) collects metadata

[Apache Jena](#) for building semantic web and Linked Data applications

[RDF4J](#) for processing and handling RDF data

[Blazegraph](#) a graph database

Contact us

- Sorbonne University: data-bsu@sorbonne-universite.fr
- University of Milan: dataverse@unimi.it
- University of Copenhagen: datamanagement@ku.dk
- Heidelberg University: data@uni-heidelberg.de
- University of Warsaw: oa.buw@uw.edu.pl
- Charles University: researchdata@cuni.cz
- University of Geneva: researchdata-info@unige.ch

4eu+

Open for you!

An introduction series to Open Science II

<https://4euplus.eu/4EU-498.html>

<https://zenodo.org/communities/4euplus-open-science/>

Open for you! An introduction series to Open Science II 26 June 2023

