CBS COPENHAGEN BUSINESS SCHOOL

KOR Registerforskningskonference 2024

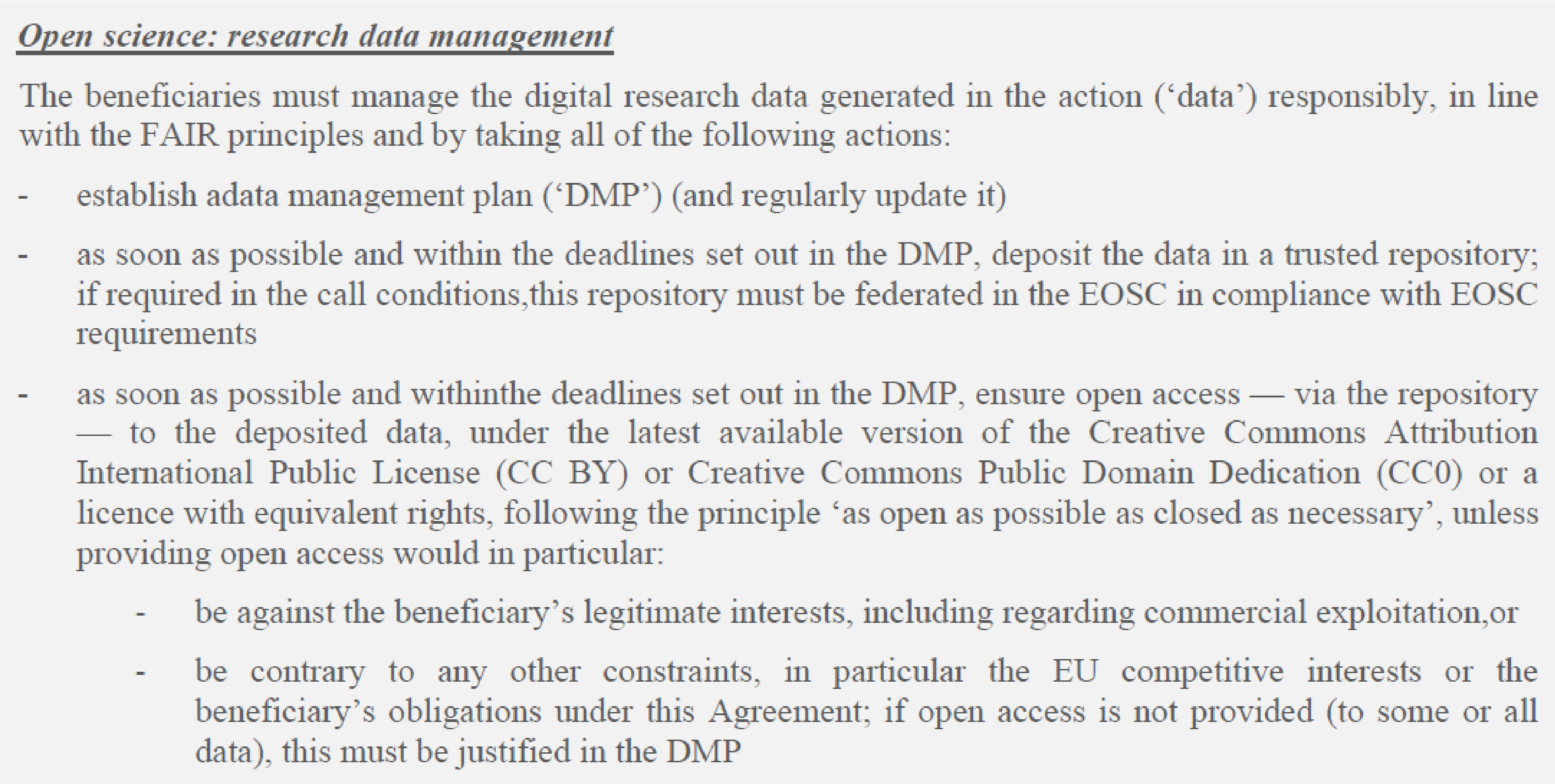
FAIR DATA

Mareike Christina Harms Buss (mabu.lib@cbs.dk)
RDM Senior Adviser, Copenhagen Business School

EQUIS ACCREDITED · AACSB ACCREDITED · AMBA ACCREDITED · CEMS · PIM

Funders require FAIR data

FAIR and responsible data management is a central funder requirement:



Open science: research data management

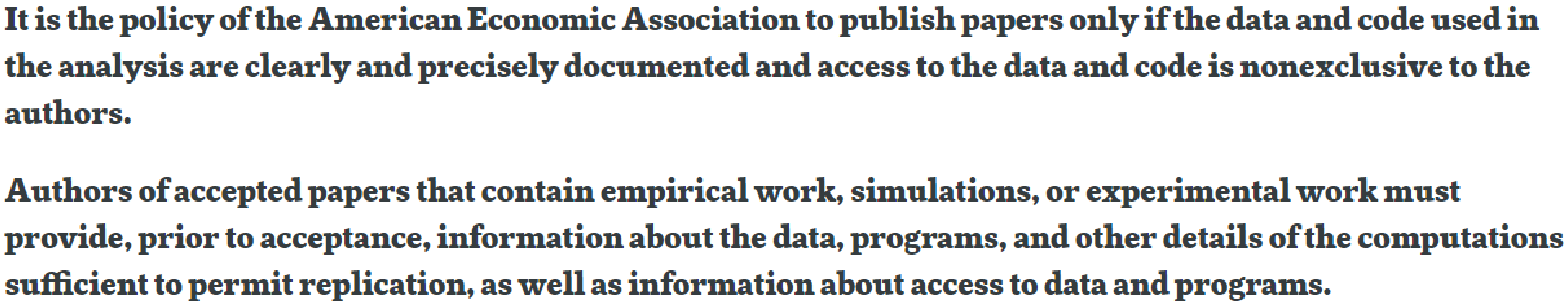
The beneficiaries must manage the digital research data generated in the action ('data') responsibly, in line with the FAIR principles and by taking all of the following actions:

- establish adata management plan ('DMP') (and regularly update it)
- as soon as possible and within the deadlines set out in the DMP, deposit the data in a trusted repository; if required in the call conditions,this repository must be federated in the EOSC in compliance with EOSC requirements
- as soon as possible and withinthe deadlines set out in the DMP, ensure open access — via the repository — to the deposited data, under the latest available version of the Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC0) or a licence with equivalent rights, following the principle 'as open as possible as closed as necessary', unless providing open access would in particular:
    - be against the beneficiary's legitimate interests, including regarding commercial exploitation,or
    - be contrary to any other constraints, in particular the EU competitive interests or the beneficiary's obligations under this Agreement; if open access is not provided (to some or all data), this must be justified in the DMP

Horizon Europe Annotated Model Grant Agreement

Journals require reproducible or replicable data

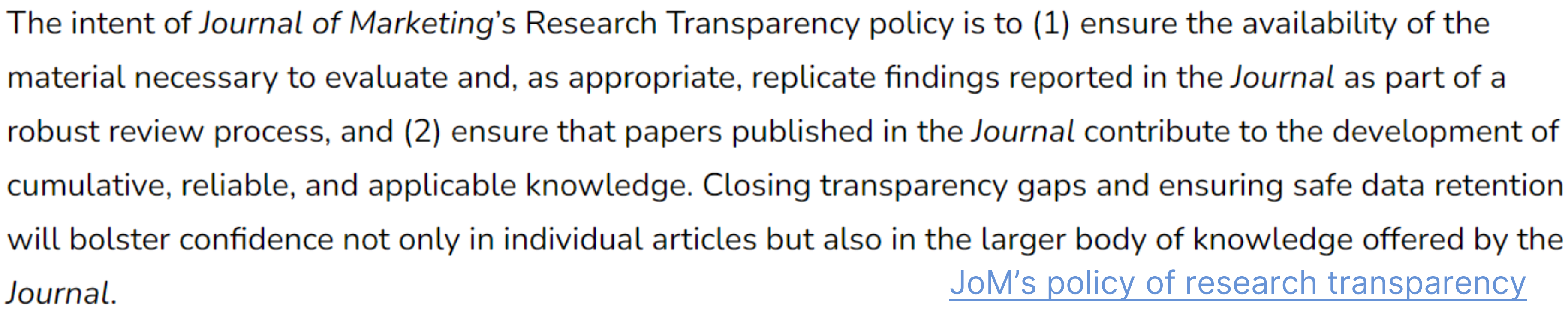
Data reproducibility or replicability is a central journal requirement:



It is the policy of the American Economic Association to publish papers only if the data and code used in the analysis are clearly and precisely documented and access to the data and code is nonexclusive to the authors.

Authors of accepted papers that contain empirical work, simulations, or experimental work must provide, prior to acceptance, information about the data, programs, and other details of the computations sufficient to permit replication, as well as information about access to data and programs.

AEA Data and Code Availability Policy



The intent of *Journal of Marketing*'s Research Transparency policy is to (1) ensure the availability of the material necessary to evaluate and, as appropriate, replicate findings reported in the *Journal* as part of a robust review process, and (2) ensure that papers published in the *Journal* contribute to the development of cumulative, reliable, and applicable knowledge. Closing transparency gaps and ensuring safe data retention will bolster confidence not only in individual articles but also in the larger body of knowledge offered by the *Journal*.

JoM's policy of research transparency

What do researchers want to do with their *primary data* after the end of the project?

- *Publish the data for reuse*, e.g. as a data paper or in a data repository:
  - FAIR data package

- *Publish the data for replication*, e.g. as supplementary material to an article:
  - Replication package
  - (Executable) research compendium

- *Not publish the data*, but curate them for own data reuse

But what about secondary data?

And, more specifically, confidential secondary data like register data?

CBS COPENHAGEN BUSINESS SCHOOL

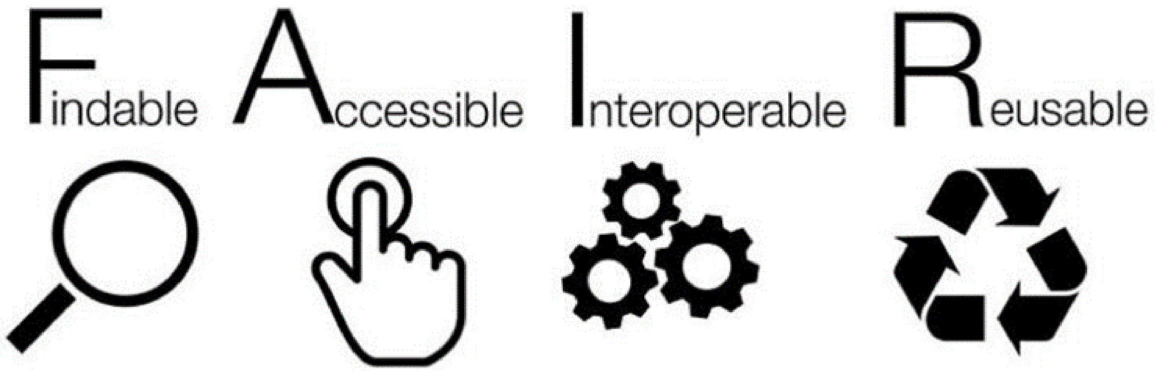
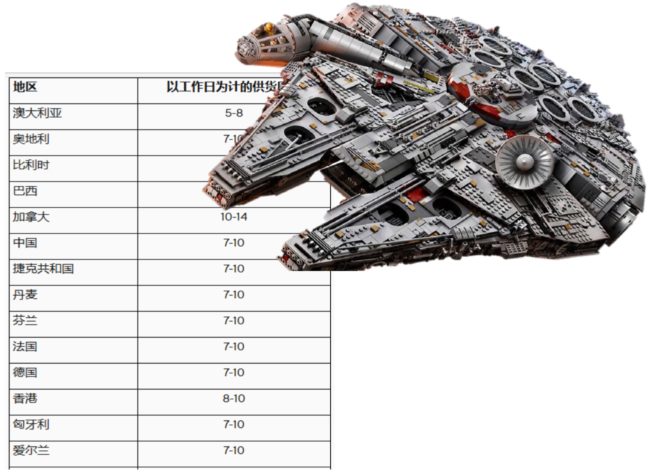


Image: CC-BY-SA 4.0 by Sangya Pundir

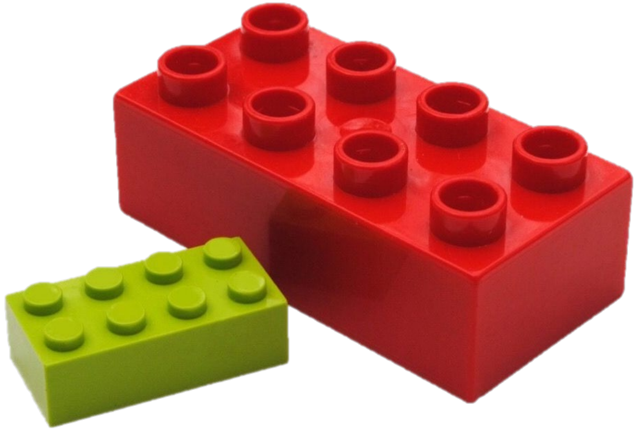
HOW TO CREATE FAIR DATA PACKAGES

The FAIR principles



地区	以工作日为计的供货
澳大利亚	5-8
奥地利	7-1
比利时	
巴西	
加拿大	10-14
中国	7-10
捷克共和国	7-10
丹麦	7-10
芬兰	7-10
法国	7-10
德国	7-10
香港	8-10
匈牙利	7-10
爱尔兰	7-10

http://www.lego.com/404



Slide by Falco Hüser, Royal Library

The FAIR principles

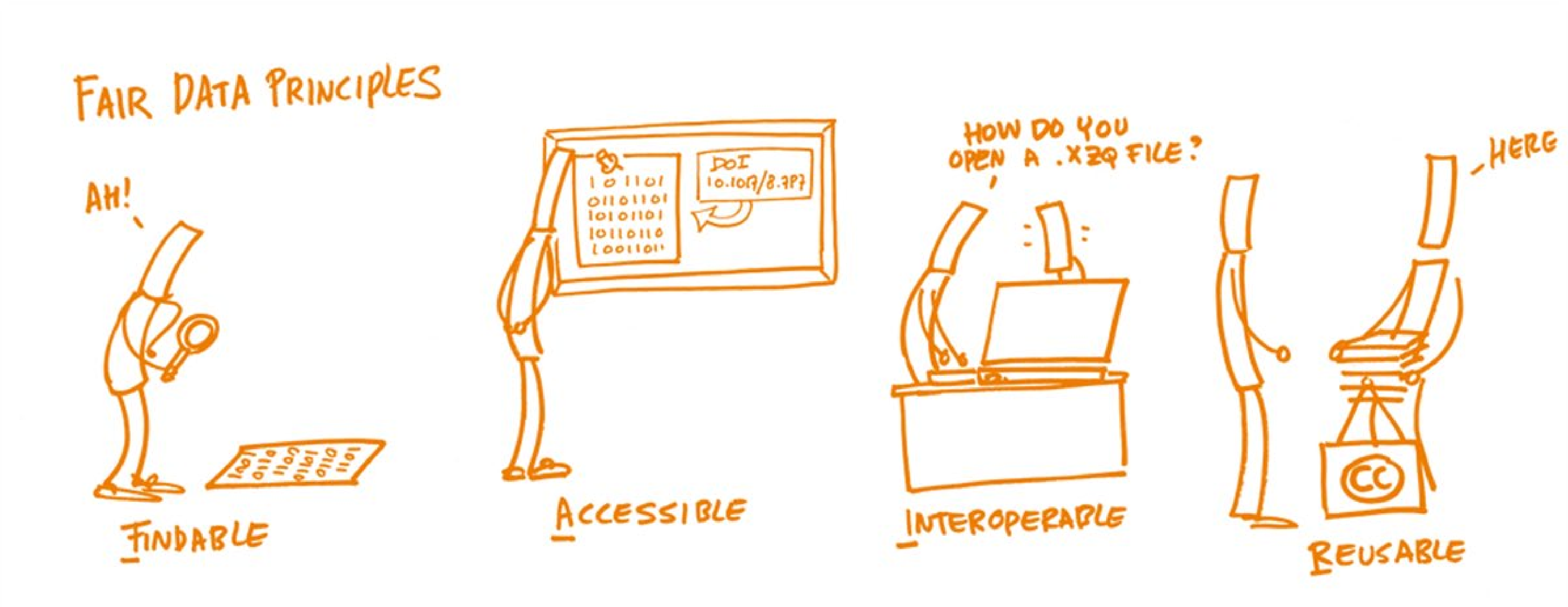


Illustration by Patrick Hochstenbach published in the Open Science Training Handbook under CC0

FAIR PRINCIPLES

Findable:

F1. (meta)data are assigned a globally unique and persistent identifier;

F2. data are described with rich metadata;

F3. metadata clearly and explicitly include the identifier of the data it describes;

F4. (meta)data are registered or indexed in a searchable resource;

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;

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;

Reusable:

R1. (meta)data are richly described with a plurality of accurate and relevant attributes;

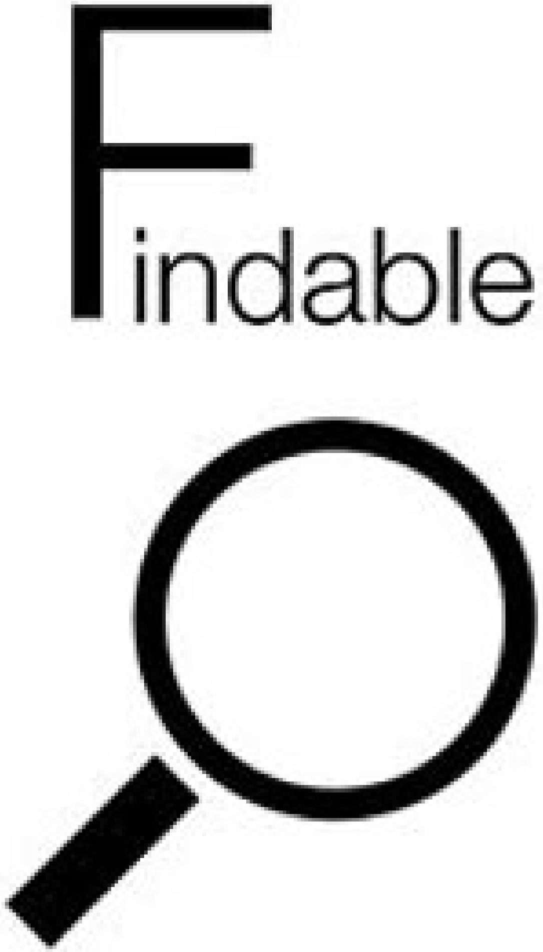
- R1.1. (meta)data are released with a clear and accessible data usage license;
- R1.2. (meta)data are associated with detailed provenance;
- R1.3. (meta)data meet domain-relevant community standards;

Making your data FINDABLE

Publish your data in a **searchable ressource** like Dataverse, Zenodo or Figshare

Making your data FINDABLE

Assign **persistent identifiers**, e.g. ORCID, DOI, ROR



Making your data FINDABLE

Provide rich metadata

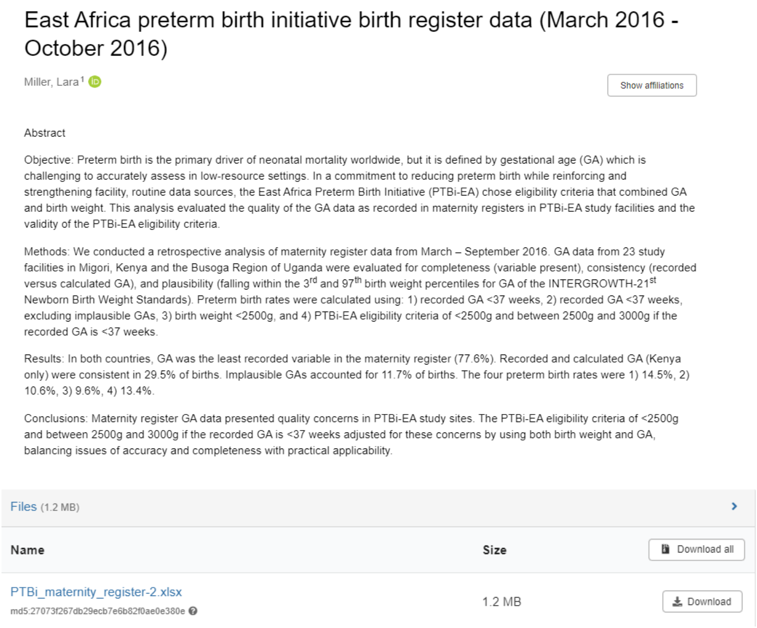
Example: Zenodo

Data set title

Author name & ORCID

Data set description

File names



East Africa preterm birth initiative birth register data (March 2016 - October 2016)

Miller, Lara[1]

Abstract

Objective: Preterm birth is the primary driver of neonatal mortality worldwide, but it is defined by gestational age (GA) which is challenging to accurately assess in low-resource settings. In a commitment to reducing preterm birth while reinforcing and strengthening facility, routine data sources, the East Africa Preterm Birth Initiative (PTBi-EA) chose eligibility criteria that combined GA and birth weight. This analysis evaluated the quality of the GA data as recorded in maternity registers in PTBi-EA study facilities and the validity of the PTBi-EA eligibility criteria.

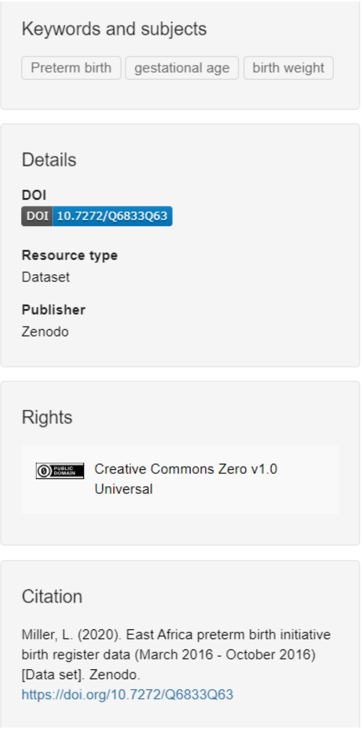
Methods: We conducted a retrospective analysis of maternity register data from March – September 2016. GA data from 23 study facilities in Migori, Kenya and the Busoga Region of Uganda were evaluated for completeness (variable present), consistency (recorded versus calculated GA), and plausibility (falling within the 3rd and 97th birth weight percentiles for GA of the INTERGROWTH-21st Newborn Birth Weight Standards). Preterm birth rates were calculated using: 1) recorded GA <37 weeks, 2) recorded GA <37 weeks, excluding implausible GAs, 3) birth weight <2500g, and 4) PTBi-EA eligibility criteria of <2500g and between 2500g and 3000g if the recorded GA is <37 weeks.

Results: In both countries, GA was the least recorded variable in the maternity register (77.6%). Recorded and calculated GA (Kenya only) were consistent in 29.5% of births. Implausible GAs accounted for 11.7% of births. The four preterm birth rates were 1) 14.5%, 2) 10.6%, 3) 9.6%, 4) 13.4%.

Conclusions: Maternity register GA data presented quality concerns in PTBi-EA study sites. The PTBi-EA eligibility criteria of <2500g and between 2500g and 3000g if the recorded GA is <37 weeks adjusted for these concerns by using both birth weight and GA, balancing issues of accuracy and completeness with practical applicability.

Files (1.2 MB)

Name	Size	Download all
PTBi_maternity_register-2.xlsx md5:27073f267db29ecb7e6b82f0ae0e380e	1.2 MB	Download



Keywords and subjects

Details

DOI

DOI 10.7272/Q6833Q63

Resource type

Dataset

Publisher

Zenodo

Rights

Creative Commons Zero v1.0 Universal

Citation

Miller, L. (2020). East Africa preterm birth initiative birth register data (March 2016 - October 2016) [Data set]. Zenodo. https://doi.org/10.7272/Q6833Q63

Keywords

Persistent identifier (DOI)

License

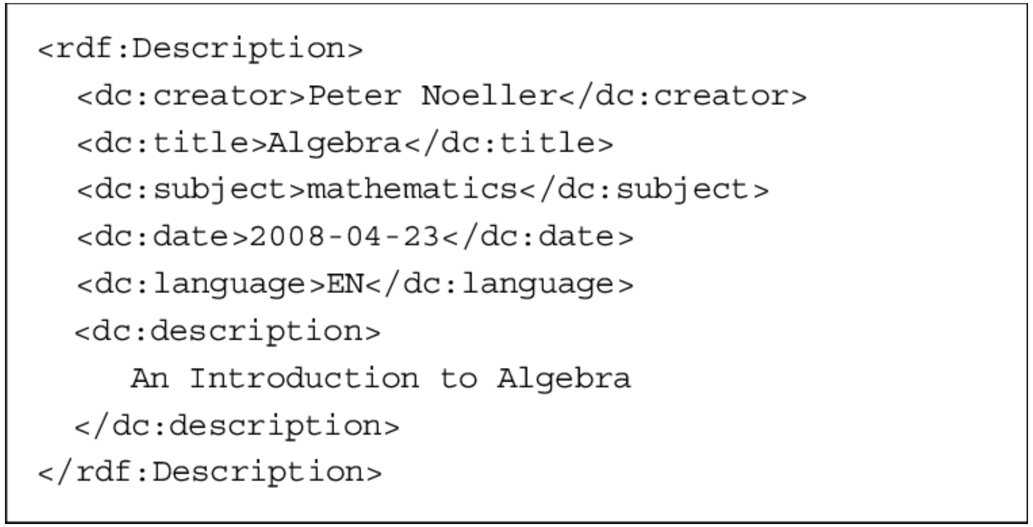
Making your data ACCESSIBLE

- Data are **retrievable by their identifier** using a standard protocol
- **Metadata are accessible**, even if the data are closed
  - ➢ Open Access
  - ➢ Embargoed Access
  - ➢ Restricted Access
  - ➢ No Access

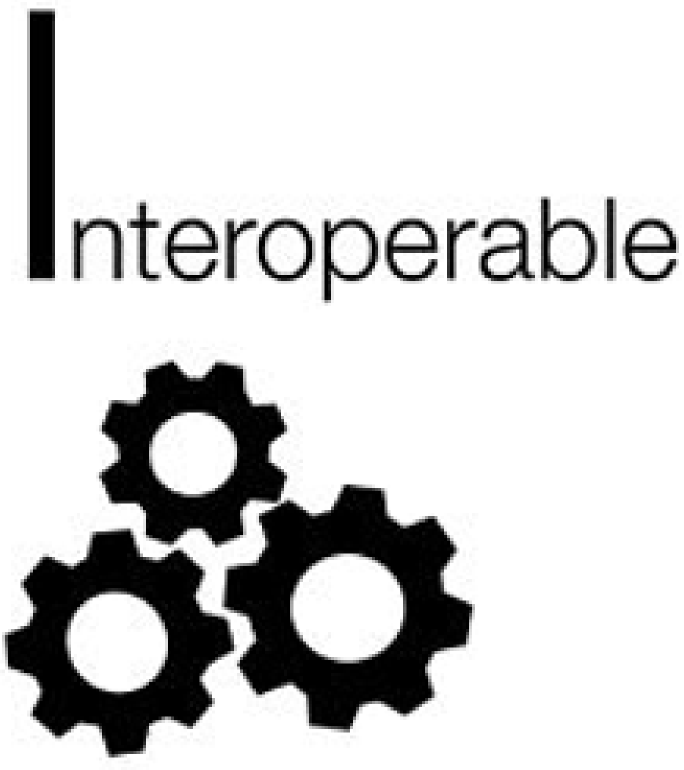
Making your data INTEROPERABLE

- Use a **formal machine-readable language** for the metadata

Examples
RDF
JSON



```
<rdf:Description>
  <dc:creator>Peter Noeller</dc:creator>
  <dc:title>Algebra</dc:title>
  <dc:subject>mathematics</dc:subject>
  <dc:date>2008-04-23</dc:date>
  <dc:language>EN</dc:language>
  <dc:description>
     An Introduction to Algebra
  </dc:description>
</rdf:Description>
```



Making your data INTEROPERABLE

- Use a **formal machine-readable language** for the metadata
- Use **open file formats** (if possible)

Examples
CSV for tabular data
RTF for textual data
TIFF for images

Making your data INTEROPERABLE

- Use a **formal machine-readable language** for the metadata
- Use **open file formats** (if possible)
- Use **common standards** (discipline)

Examples

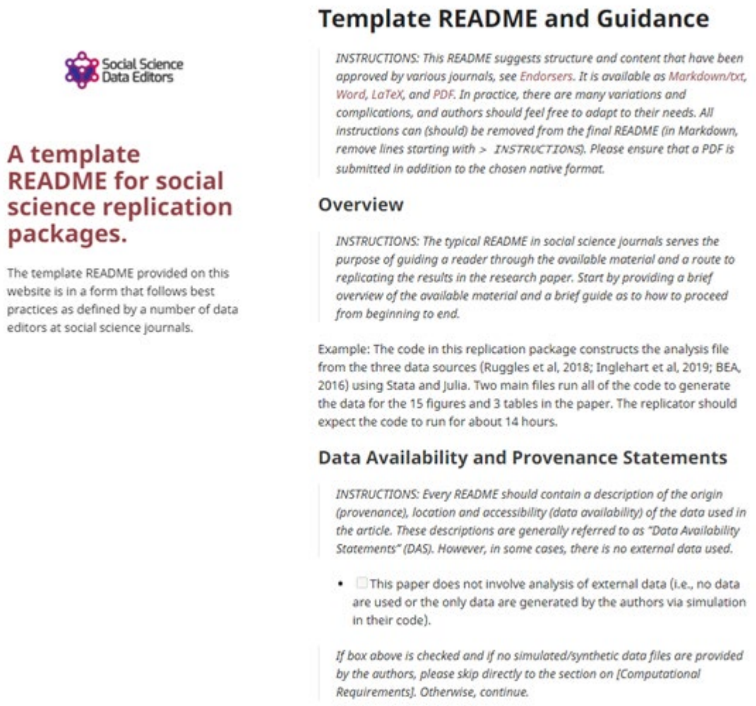
➢ [Data and Code Availability Standard v1.0]

DCAS

Data and Code Availability Standard

Making your data REUSABLE

- Provide **rich accurate metadata**
- Give **detailed provenance information**
- Refer to **community standards**



Social Science Data Editors

A template README for social science replication packages.

The template README provided on this website is in a form that follows best practices as defined by a number of data editors at social science journals.

Template README and Guidance

> *INSTRUCTIONS: This README suggests structure and content that have been approved by various journals, see Endorsers. It is available as Markdown/txt, Word, LaTeX, and PDF. In practice, there are many variations and complications, and authors should feel free to adapt to their needs. All instructions can (should) be removed from the final README (in Markdown, remove lines starting with `> INSTRUCTIONS`). Please ensure that a PDF is submitted in addition to the chosen native format.*

Overview

> *INSTRUCTIONS: The typical README in social science journals serves the purpose of guiding a reader through the available material and a route to replicating the results in the research paper. Start by providing a brief overview of the available material and a brief guide as to how to proceed from beginning to end.*

Example: The code in this replication package constructs the analysis file from the three data sources (Ruggles et al, 2018; Inglehart et al, 2019; BEA, 2016) using Stata and Julia. Two main files run all of the code to generate the data for the 15 figures and 3 tables in the paper. The replicator should expect the code to run for about 14 hours.

Data Availability and Provenance Statements

> *INSTRUCTIONS: Every README should contain a description of the origin (provenance), location and accessibility (data availability) of the data used in the article. These descriptions are generally referred to as "Data Availability Statements" (DAS). However, in some cases, there is no external data used.*

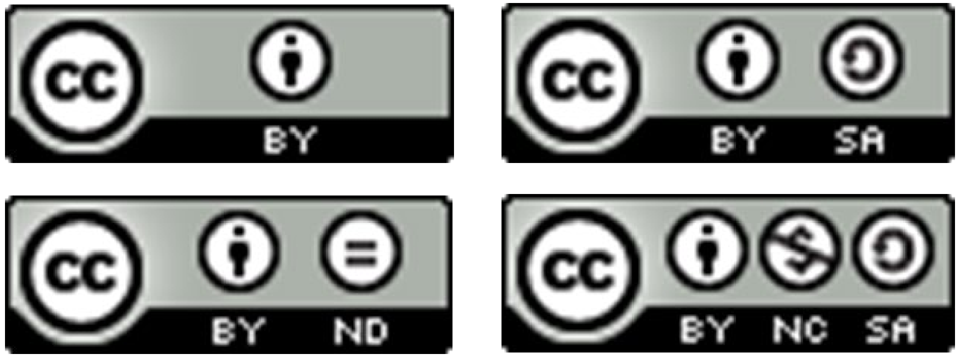
- □ This paper does not involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).

> *If box above is checked and if no simulated/synthetic data files are provided by the authors, please skip directly to the section on [Computational Requirements]. Otherwise, continue.*

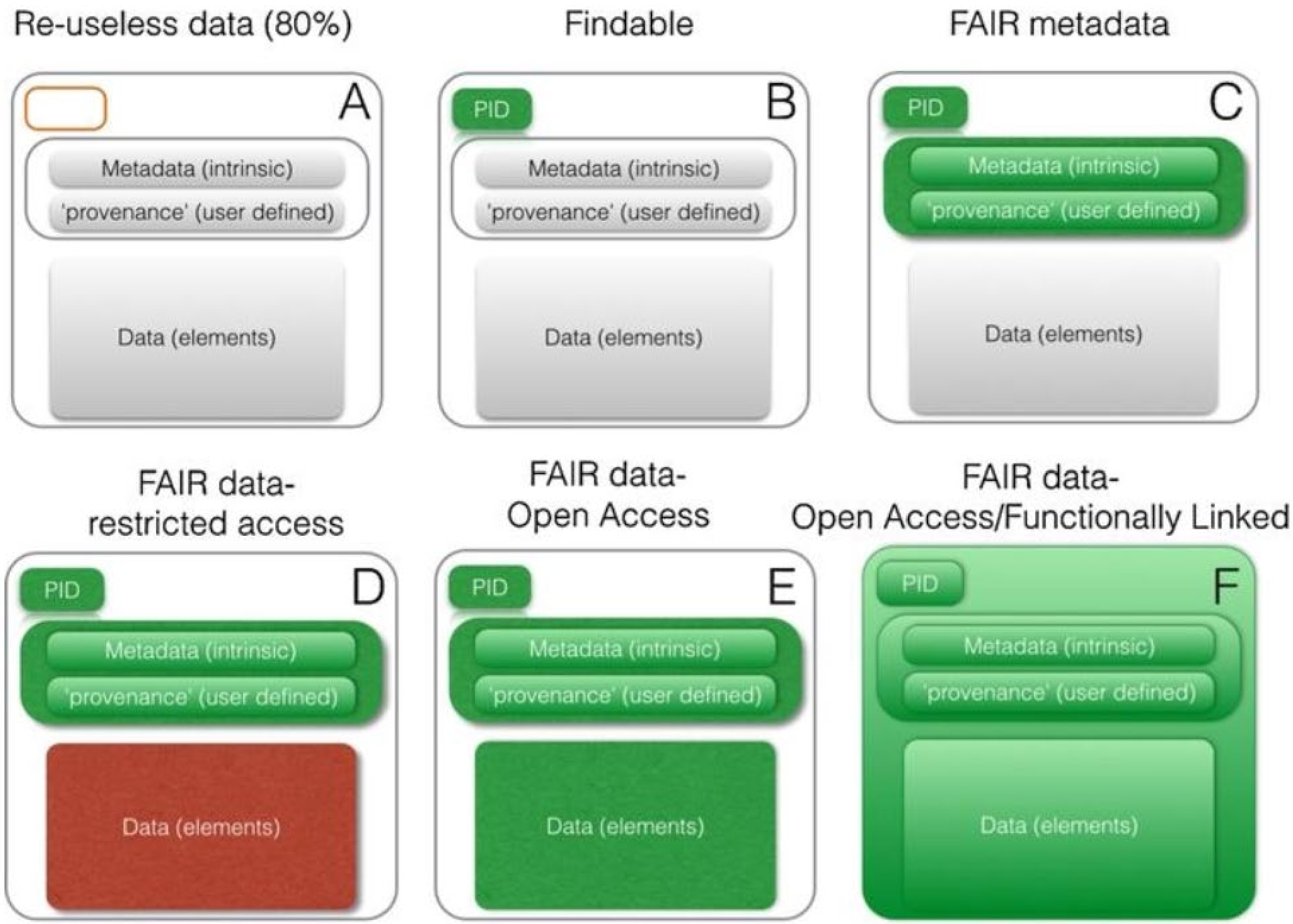
https://social-science-data-editors.github.io/template_README/template-README.html

Making your data REUSABLE

- Provide **rich accurate metadata**
- Give **detailed provenance information**
- Refer to **community standards**
- Use **clear licenses** for reuse, e.g. creative commons licenses

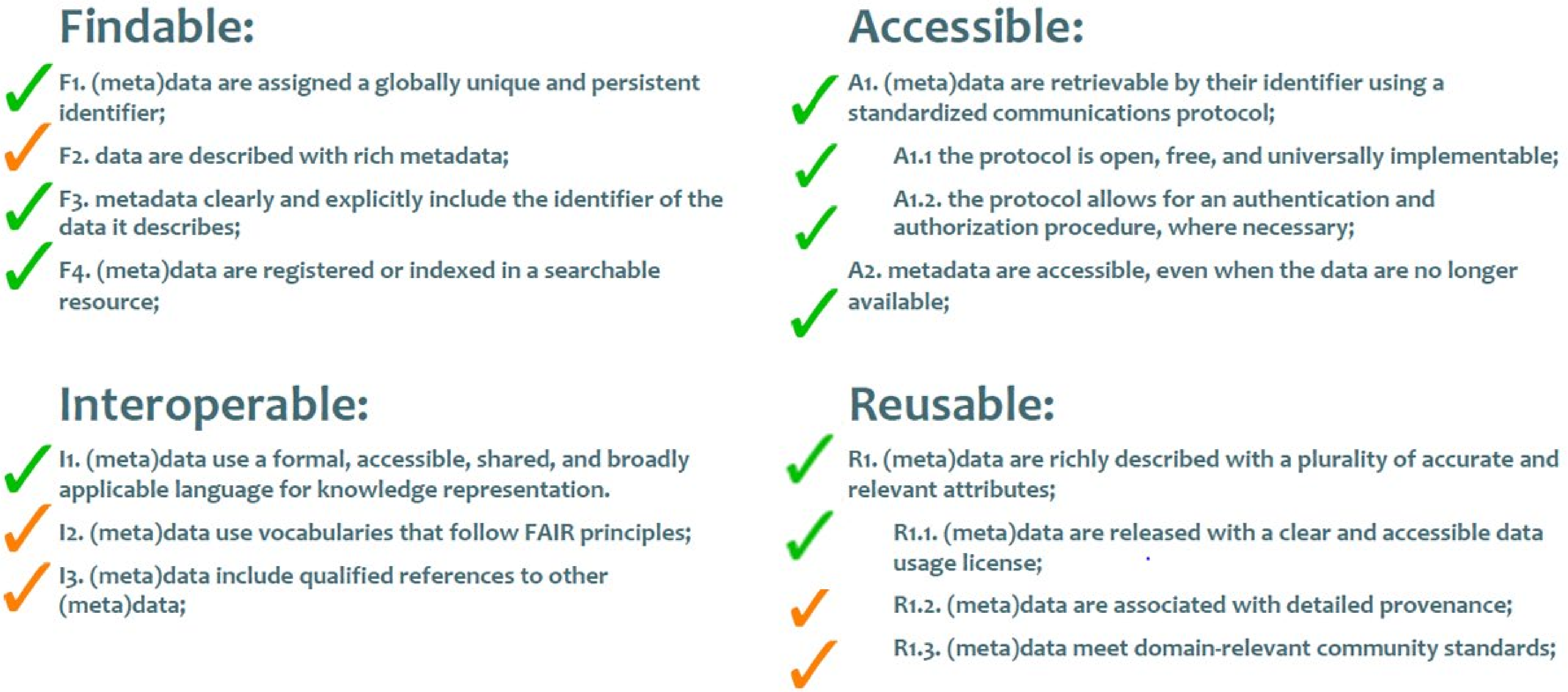


Shades of FAIR



Mons, B. et al. 2017. ”Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud.” Information Services & Use 37/1:49-56. DOI: 10.3233/ISU-170824

Repositories support data FAIRification



Findable:

- ✓ F1. (meta)data are assigned a globally unique and persistent identifier;
- ✓ F2. data are described with rich metadata;
- ✓ F3. metadata clearly and explicitly include the identifier of the data it describes;
- ✓ F4. (meta)data are registered or indexed in a searchable resource;

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;

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;

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;

Register data are FAIR by default

Findable

- Rich, standardized metadata
- Variable lists

BEF – https://www.dst.dk/extranet/ForskningVariabellister/BEF%20-%20Befolkningen.html

DOD – https://www.dst.dk/extranet/ForskningVariabellister/DOD%20-%20D%C3%B8de%20i%20Danmark.html

FAIN – https://www.dst.dk/extranet/ForskningVariabellister/FAIN%20-%20Husstande%20og%20familier.html

Accessible

- Accessible (under an authorization)

Access to data

Access to data under the Research Scheme

Interoperable

- With other register data
- With external data, if linked by key variable

Reusable

- Reusability is at the core of DST

STATISTICS DENMARK

CBS

Burmeister, N. B. (2022). Arbejdsgruppe A - Use Case – FAIR Access to Register Data. Zenodo. https://doi.org/10.5281/zenodo.7437332

Register data are FAIR by default – with limitations

Findable

- Easily findable - for experts only
- You don’t always find what you need

Accessible

- Only Danish institutions can be authorized
- (Language barrier)

Interoperable

- Limited interoperability with other register data from the Nordics

Reusable

- Changing variables (provenance!)
- Difficult access to previous projects

Burmeister, N. B. (2022). Arbejdsgruppe A - Use Case – FAIR Access to Register Data. Zenodo. https://doi.org/10.5281/zenodo.7437332

CBS COPENHAGEN BUSINESS SCHOOL

HOW TO CREATE REPLICABLE DATA PACKAGES

EQUIS ACCREDITED AACSB ACCREDITED AMBA ACCREDITED CEMS PIM

Reproducibility vs. replicability

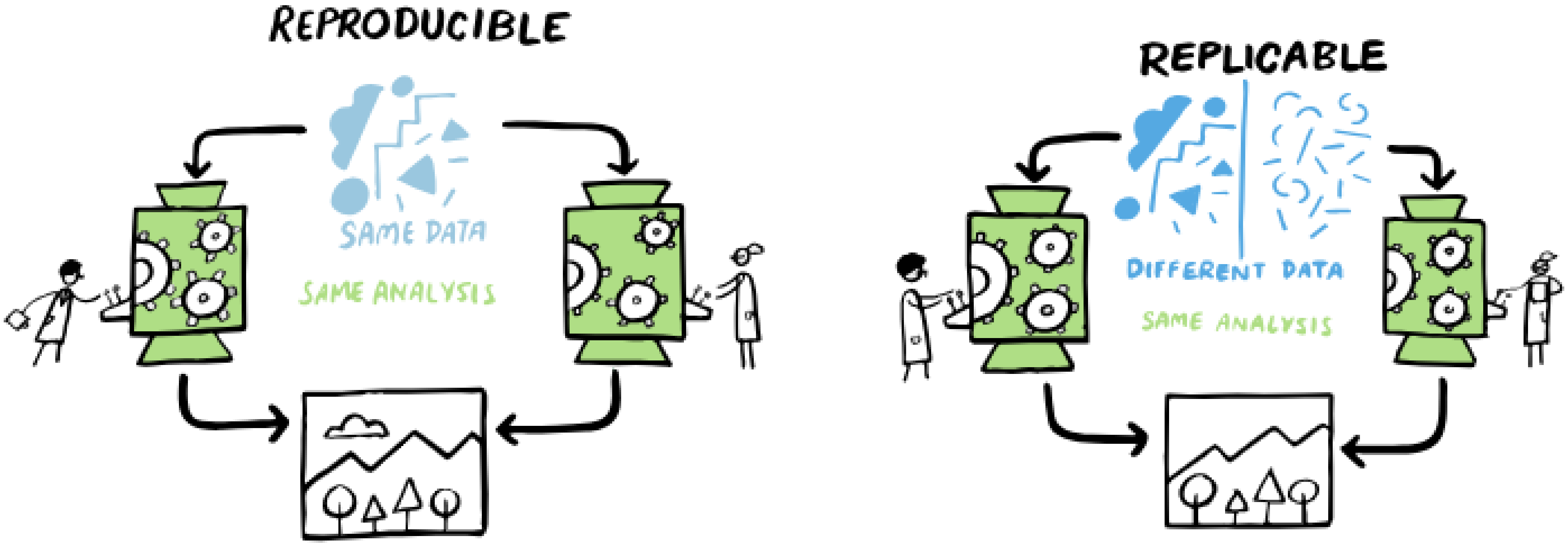


Image: The Turing Way Community, & Scriberia. (2024). *Illustrations from The Turing Way*: Shared under CC-BY 4.0 for reuse. Zenodo. https://doi.org/10.5281/zenodo.13882307

DCAS – the gold standard for replication packages

Data and Code Availability Standard (DCAS) requires the following elements:

1. Data availability statement
2. Raw data
3. Analysis data
4. Format
5. Metadata
6. Citation

DCAS – the gold standard for replication packages

Data and Code Availability Standard (DCAS) requires the following elements:

7. Data transformation
8. Analysis
9. Format

DCAS – the gold standard for replication packages

Data and Code Availability Standard (DCAS) requires the following elements:

10. Instruments
11. Ethics
12. Pre-registration
13. Documentation
14. Location
15. License
16. Omissions

How to comply with DCAS when working with register data?

Create a package consisting of:

- Data citations
- Data access description
- Code
- Supplementary material

FAIR protocol

Squaring the circle: FAIR and replicable register data packages

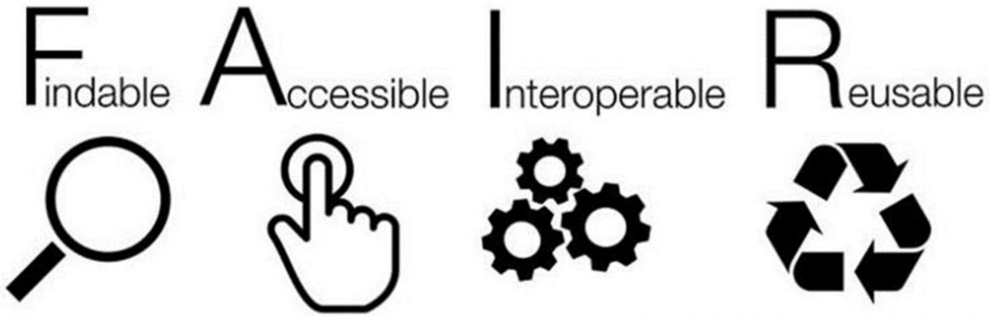


Image: CC-BY-SA 4.0 by Sangya Pundir

DCAS

Data and Code Availability Standard