Eötvös Loránd University
Faculty of Education and Psychology

FAIR Data Management

Dr. Tamás Nagy
ELTE Eötvös Loránd University
@nagyt

Introduction

Tamás Nagy, Assistant professor at ELTE Eötvös Loránd University.

Original training in Psychology.

Engaged with data science, and data analysis issues and open science.

Overview of the presentation

1. FAIR data principles and meaning.
2. Good practices for sharing FAIR data.

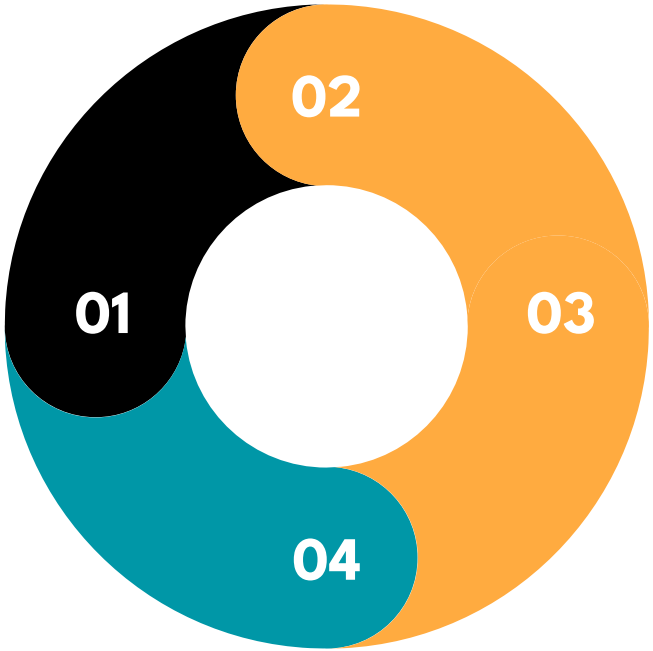
Why share data?

- Universism: prove by evidence and not by authority (Merton, 1942).
- This also means that everyone has to be able to verify the evidence, based on data.
- By that time paper was the main information carrier, thus data sharing was cumbersome.
- Later came the computers, digital data, and the internet.
- Science was still stuck in paper-based information sharing until recently.
- Scientists should embrace the possibility to share data (as it was originally intended).
- Sharing data facilitates good practices in data management.

FAIR data

Findable

Accessible



Reusable

Interoperable

Findable Data

- Data are findable if it can be easily located and accessed by users or systems.
- This means that it is **indexed** relevant by search engines.
- **Persistent identifiers** (PIDs) — like DOI — are keys for indexing, as it means that independent of the storage space.
- Several sites offer generating a DOI for free (e.g. osf.io, zenodo.org).
- Some journals publish datasets (e.g., Journal of Open Psychology Data).

Accessible Data

- Data should be accessible to **everyone** at any time.
- This means that **no burdens** are posed by the owner (or storage space) to get the data.
- It also means that the storage space is maintained.
- Scientific data **repositories** (e.g. osf.io) make it easy to store and share data, and provide long-term maintenance.

Interoperable Data

- Interoperability means that data is usable **irrespective of the software or operating** system of the user.
- Data should be in an **open format**, i.e., users should be able to use the data with any software (e.g., csv, xml, json, API endpoint, etc.).
- Proprietary software often use their own data format which limits accessibility.
- There are also **metadata standards**, that are ensuring that data is human and machine readable (e.g., Dublin Core).

Reusable Data

- The ability of data to be easily reused for different purposes **beyond the original research context** (e.g., reproduction, meta-analysis, mega-analysis, etc.).
- Data should be well-documented, structured, and prepared in a way that allows **other researchers** to understand, interpret, and effectively utilize the data.
- It is a good practice to make the **language of the dataset** and documentation accessible for those not speaking the local language by providing translations.
- Data licencing.

Data dictionary/ Code book

1. Information about the **variables** in the dataset.
2. Information about the **data processing.**
3. Information about the **data collection procedure**.

This data frame contains 456 observations (rows), each representing a movie, and 27 variables (columns):

1. `title`: Title of movie
2. `audience_score`: Audience score on Rotten Tomatoes (response variable)
3. `type`: Type of movie (Documentary, Feature Film, TV Movie)
4. `genre`: Genre of movie (Action & Adventure, Comedy, Documentary, Drama, Horror, Mystery & Suspense, Other)
5. `runtime`: Runtime of movie (in minutes)
6. `year`: Year the movie is released
7. `mpaa_rating`: MPAA rating of the movie (G, PG, PG-13, R, Unrated)

Licences

- Sharing does not mean giving up all rights!
- We can use licenses to define what we do not allow.
- E.g., data can be freely accessed, copied, analysed, but publication from it may require a permission.
- If we do not use our own data, we must be able to prove that we have rights to publish the results from the data.
- https://choosealicense.com/

Elements of Creative commons licences

Attribute ownership: the user must indicate the source and if any changes have been made.

CC-BY 4.0

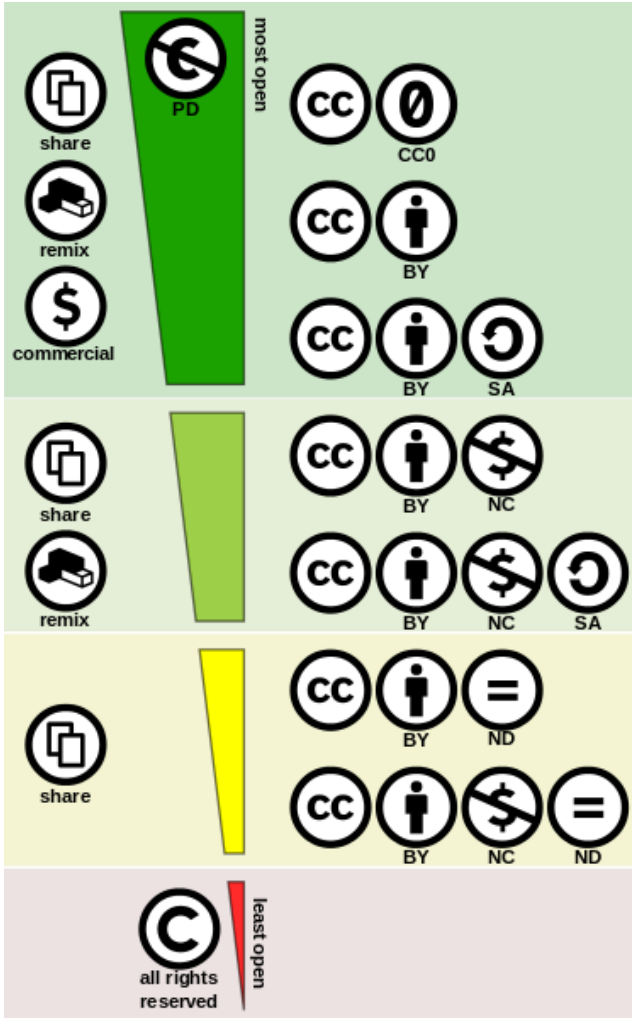
No derivatives: the user cannot modify the material or publish ne material based on it.

CC-BY-ND 4.0

Share alike: material based on it can only be published with the same license as the original.

CC-BY-SA 4.0

No commercial use: the material may not be used for commercial


most open
PD
share
remix
commercial
CC0
BY
BY
SA
share
remix
BY
NC
BY
NC
SA
share
BY
ND
BY
NC
ND
all rights
reserved
least open

Best practices for sharing research data

- Digitally shareable data, accessible to everyone.
- In a repository with a **permanent identifier** and **time stamp** (e.g., OSF).
- Includes a **codebook** to help other researchers understand the structure of the data and the content of the variables.
- A **license** that specifies that others can copy and distribute the data.
- + Compliance with data management principles and rules (anonymisation, transparency to participants).
- Open Data is linked in the published journal article.

What to share?

1. Raw data
    - Data generated at the time of recording, unchanged, without modification, deletion or aggregation.
2. And/or aggregated data
    - Square format: each observation is one row, each variable one column.
    - Each observation unit is assigned a unique identifier.
3. Code book on variables
    - Information on each variable (including unit of measurement).
    - Information on how the variables are calculated (e.g. aggregates).
    - Information on the design of the study and how the data were collected.
4. Precise description of how the data were transferred from the raw form to the processed form (1->2), if possible with program code.

Ellis, S. E., & Leek, J. T. (2018). How to share data for collaboration. *The American Statistician*, *72*(1), 53–57. https://doi.org/10.1080/00031305.2017.1375987

Example sentences for data sharing

Type of sharing	Example sentence
Data sharing part in the IRB submission	*The data collected during the research will be published in a non-personally identifiable, anonymised form at [repository name].*
Research data and materials are available on a public repository.	*All data and materials have been made publicly available at the [repository name] and can be accessed at [persistent URL or DOI].*
Anonymised research data and materials are available on public repositories.	*Anonymized data and materials have been made publicly available at the [repository name] and can be accessed at [persistent URL or DOI].*
The data of the analysis are available on a public repository.	*The data used for the analyses have been made publicly available at the [repository name] and can be accessed at [persistent URL or DOI].*

Recap

When sharing data, make sure that the data are

- Findible
- Accessible
- Interoperable
- Reusable

Eötvös Loránd University
Faculty of Education and Psychology

Thank you for your attention!

Dr. Tamás Nagy
ELTE Eötvös Loránd University
@nagyt