Skills 4 eosc

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

INFN

CNAF, LNGS, LNF, Torino

The role of the Data Steward at INFN

... and other tales of FAIR

Luca dell'Agnello, **Stefano Dal Pra**,
Luciano Gaido, Francesca Marchegiani,
Irene Piergentili, **Lorenzo Rinaldi**

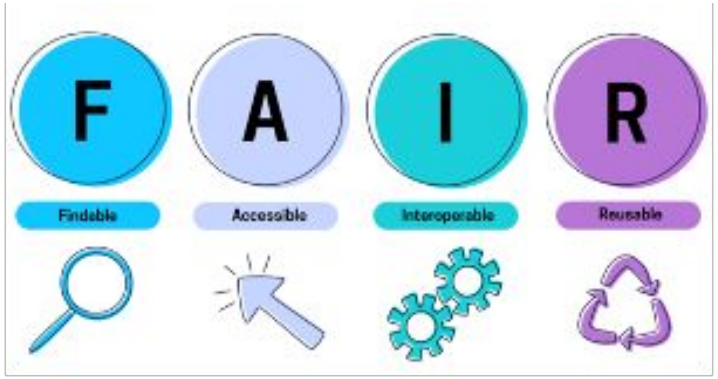
OAR link: 10.15161/oar.it/at3ca-ahs10

Outline

- Introduction
  - FAIR principles in High Energy and Nuclear Physics
- The INFN and the role of Data Stewards
  - People and connections
- A Decision Tree for Data Management Plan
- The new Open Access Repository
- Open Licences
- Outlook and Conclusions

Introduction: FAIR principles

- Ensure:
  - the reproducibility, transparency, and integrity of research
  - the validity of scientific results (authentic, complete, and reliable)
  - the traceability and future reuse of data
- Optimize the use of resources in case the same research is replicated
- Meet the requirements of funding agencies and data protection regulations
- Agree on data ownership and sharing
- Avoid data loss (lack of adequate documentation for their interpretation, obsolescence of formats and software that ensure their accessibility, visualization, and analysis)
- Encourage collaboration among researchers



FAIR principles in High Energy and Nuclear Physics

Already a good practice in many HENP communities

Large experiments adopted FAIR principles for:
- Data Management Plan
    - FAIR access to data and software
- Open Access policies

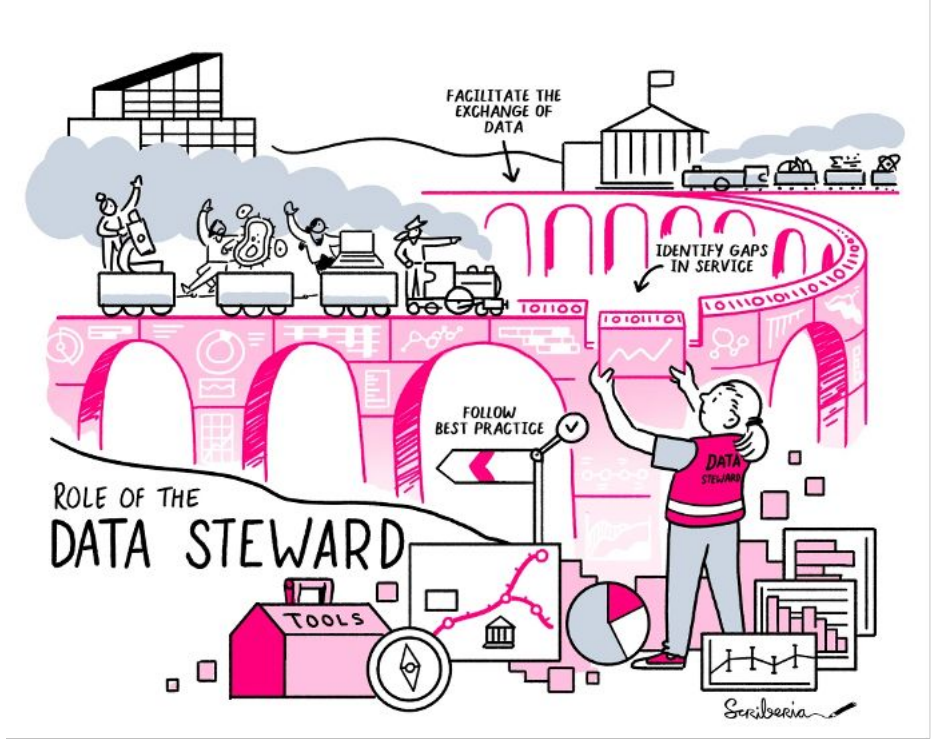
Many leading institutions for OpenScience (CERN, GSI, …)

How can small research communities or individual experiments be supported?

The role of Data Steward

Data Steward represents a new professional profile that combines expertise in data management with deep knowledge of specific scientific fields:

- in-depth knowledge in specific research areas
- responsible for the correct and effective FAIR management of research data throughout their entire lifecycle
- Disciplinary and transversal skills, team-work, experienced in Open Science topics.
- Give support for research data management (administrative and scientific-technological)



https://openworking.files.wordpress.com/2022/04/data-stewards.jpg?w=1024

Why a Data Steward Team at INFN?

Characterization of data in HENP

Large experiments vs small communities and individuals
There is no "standard" configuration
each scientific collaboration has a different approach to the various phases of data lifecycle management, depending on the choices made within their own collaboration.

Experiments funded by many international bodies

Data ownership
different funding agencies and different scientific communities

Data distribution among different entities
(multiple geographically distributed copies)

Levels of data processing (raw data, calibration data, reconstructed (pre-analyzed) data, reduced data, published data, etc.)

Data format and typologies: each experiment has its own format

"Proprietary" software: acquisition, processing, and reading software developed within each scientific collaboration to meet their own purposes and needs

Duration of experiments: 5 – 20 years

Avoiding technological obsolescence: necessary updating of routines and software with current systems to ensure compatibility with operating systems.

Open Science @ INFN

Since 2021: Institution of the **INFN OpenScience** working group
- https://web.infn.it/openscience/
- https://www.openaccessrepository.it/
- disciplinary code for open access to research products DOI: 10.15161/oar.it/211742

Collaboration and involvement in several national and European OS initiatives:
- Co-coordination of CoPER Open Science WG
- Member of the Italian Computing and Data Infrastructure (ICDI)
- Participation in EOSC projects

Italian Data Steward Community (since 2023)
Competence center for Open Science, FAIR e EOSC (CC-ICDI)

Sharing expertise and experience with the other supporters within the Skill4EOSC User Support Network

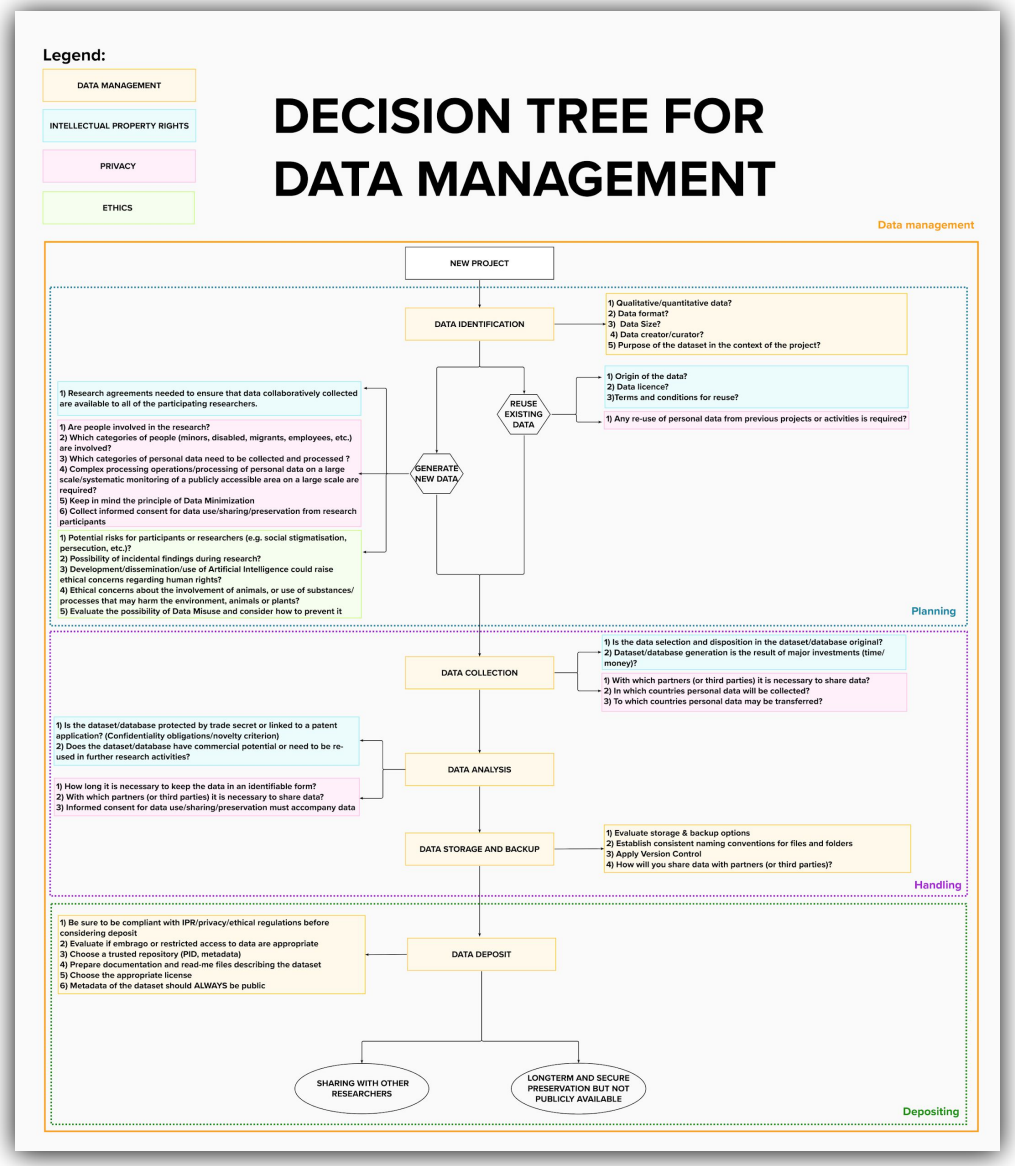
The Decision Tree for Data Management Plan

Planning
- Data identification

Handling
- Data collection
- Data analysis
- Data storage and backup

Depositing
- Data deposit:
  -Sharing
  - Long-term and secure preservation

DMP drafting in the context of HENP



Caldoni, G., Gualandi, B., & Marino, M. (2022). Research Data Management Decision Tree. Zenodo. https://doi.org/10.5281/zenodo.7190005

Planning: data identification

1) Qualitative/quantitative data—> Qualitative (conditions data) & quantitative (Physics quantities measurements), how many levels of data processing?
2) Data format —> CSV, root, json, …
3) Data Size—> from MB to TB (PB and EB are typical of large collaborations…)
4) Data creator/curator—> Small group (also international) or single researcher
5) Purpose of the dataset in the context of the project?—> Detector/auxiliary data, Physics data

Privacy / Confidentiality:

- Experiments involving people (bio-physics, nuclear medicine)
- Personal data treatment

Ethics:

- Potential risk (from radiation sources, radiation activation)
- Data misuse

Handling: data collection and data analysis

1) Data coming from an experimental facility (a small accelerator in an external lab, set an innovative detector close to a nuclear reactor, etc)
2) National or international grants…
3) Collaboration with private companies, Technology Transfer
4) International collaboration (same rules?)—> research agreement

Handling: data storage and backup

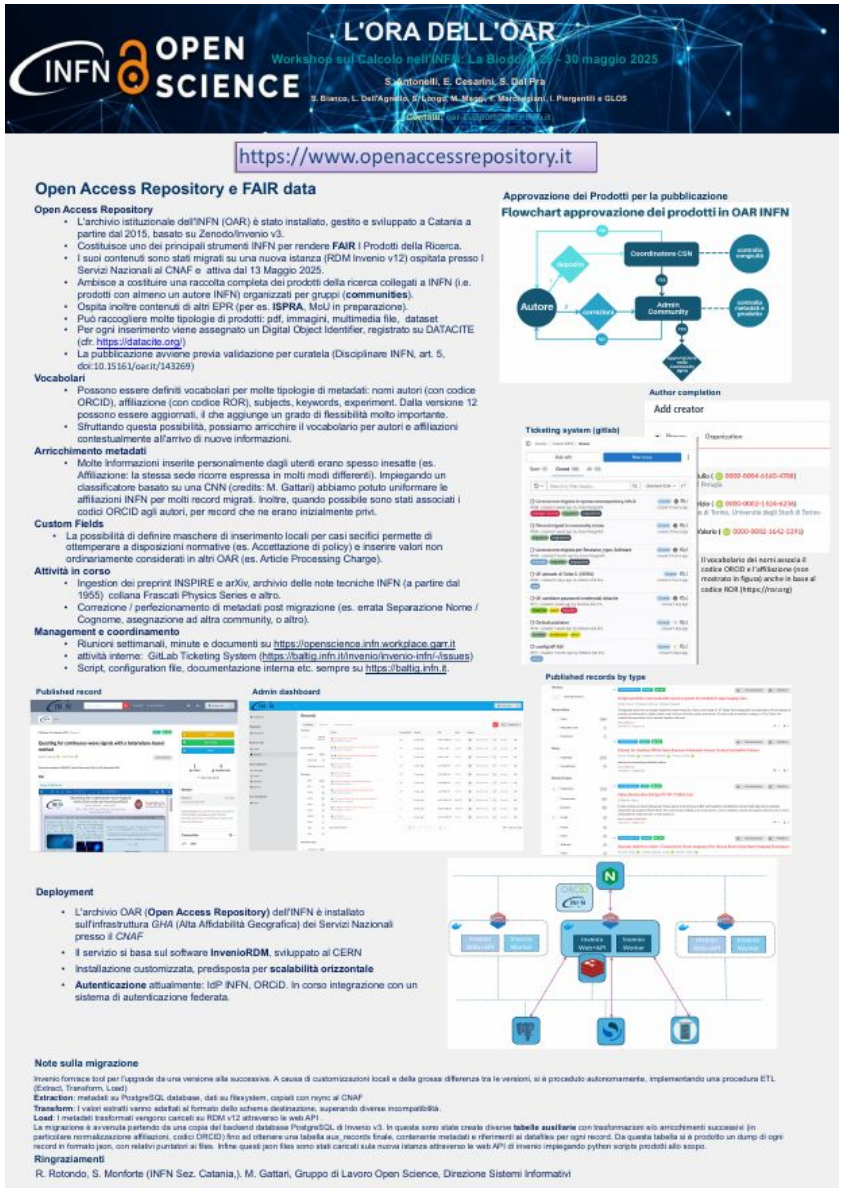
1) Many solutions available (better than private local HD…):
   - **INFN-Cloud**
   - Grid-like storage endpoints (https/webdav, xrootd access)
   - Commercial Cloud (GDrive or MS OneDrive)
2) Use of high level data management tool (Rucio…)
3) For the Software: Git-Github, Gitlab, **Baltig**
4) remote&secure access (IAM or other secure authentication), Public storage endpoint via WEB interface (for OpenData)

Depositing: the INFN Open Access Repository

https://doi.org/10.15161/oar.it/zqys6-rtz84

- Contents published upon validation (Curatela)
- DOI registration (Findability)
- Ingestion from 3rd party repos (In progress)
- Several content types, (talk, paper, datasets,...)
- Communities

- Infrastructure managed by SSNN at CNAF
  - Geographic High Availability
- OAuth: INFN IdP, ORCID; federated Oauth (Coming soon)



Open Licenses

Guidelines to be published in a next Disciplinare following Recommendations to be finalized from INFN Tech Transfer Service dedicated WG. Latest (Old) recommendations from TT are available here.
briefly:

- (Old) Recommendation from Francesco Giacomini, Lorenza Paolucci to recommend EUPL for SW products
  - Designed for European Public Institutions, to be compatible with EU legislations
  - Official legal translation on EU languages
  - Matrix Compatibility vs other licences
  - Latest version (1.2, 2017) updates against law changes, more precise definitions for Patents.
- Most popular open licecenses being considered.
- Current agreements and experiment's decisions to be considered. Dual licensing possible in some cases.

Real case:

- A research group writes a MATLAB SW Library. Mathworks wants to adopt that library in a Toolbox demo. What License do you release your library with?
  - Mathworks proposal: BSD like license
- However, with BSD:
  - No copyleft: derived work can be closed, and used into proprietary products

Outlook and conclusions

FAIR principles are already good practice in High Energy and Nuclear Physics

Small communities may need support

The role of the Data Steward is becoming increasingly important

INFN may benefit

Work plan is on the table:

- Setup of an operative documentation for supporting researchers
- Interaction with a large network of DS and research support teams
- promote FAIR initiatives (use of OAR, open licenses)

Get in touch with real use cases to implement:

data-steward-infn@lists.infn.it (not active yet!)

Skills 4 eosc

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

INFN

THANK YOU!

Stefano Dal Pra
INFN CNAF
stefano.daplra@cnaf.infn.it

Lorenzo Rinaldi
Bologna University & INFN
lorenzo.rinaldi@unibo.it

www.unibo.it

Skills 4 eosc

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

INFN

Backup

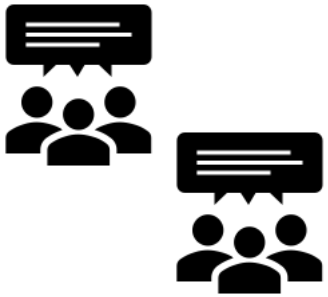
User support plan

The main objective of Data Stewards is to support small research groups
- Data Management Plan drafting, according to a precise check-list
- publication (which Open Access level?)

Target:

Researchers with no or few knowledge of FAIR principles

Different levels of support



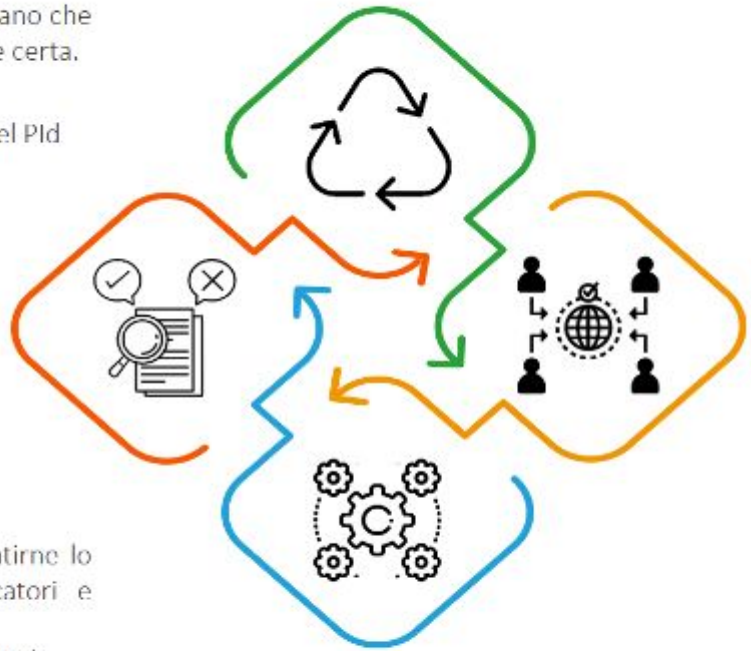
Researchers (highly) familiar with FAIR principles

What's FAIR?

Findable

Dati rintracciabili sia per l'occhio umano che per le macchine in maniera univoca e certa.

- Identificativo persistente (PId)
- Metadati descrittivi comprensivi del PId
- Ricercabili online
- Metadati indicizzati



Accessible

Dati recuperabili online attraverso protocolli standardizzati, reperibili e preservati in un orizzonte temporale a lungo termine.

- Interrogabili online con l'utilizzo di protocolli standardizzati
- Accesso limitato ai dati solo se necessario, accesso aperto ai metadati descrittivi: **As open as possible, as closed as necessary**
- Deposito in un trusted repository (es. /enodo)

Interoperable

Dati strutturati in maniera da garantirne lo scambio ed il riutilizzo tra ricercatori e istituzioni di tutto il mondo.

- Formati largamente diffusi e standards
- Vocabolari controllati
- Schemi condivisi, ontologie, parole chiave
- Evitare formati e software proprietari

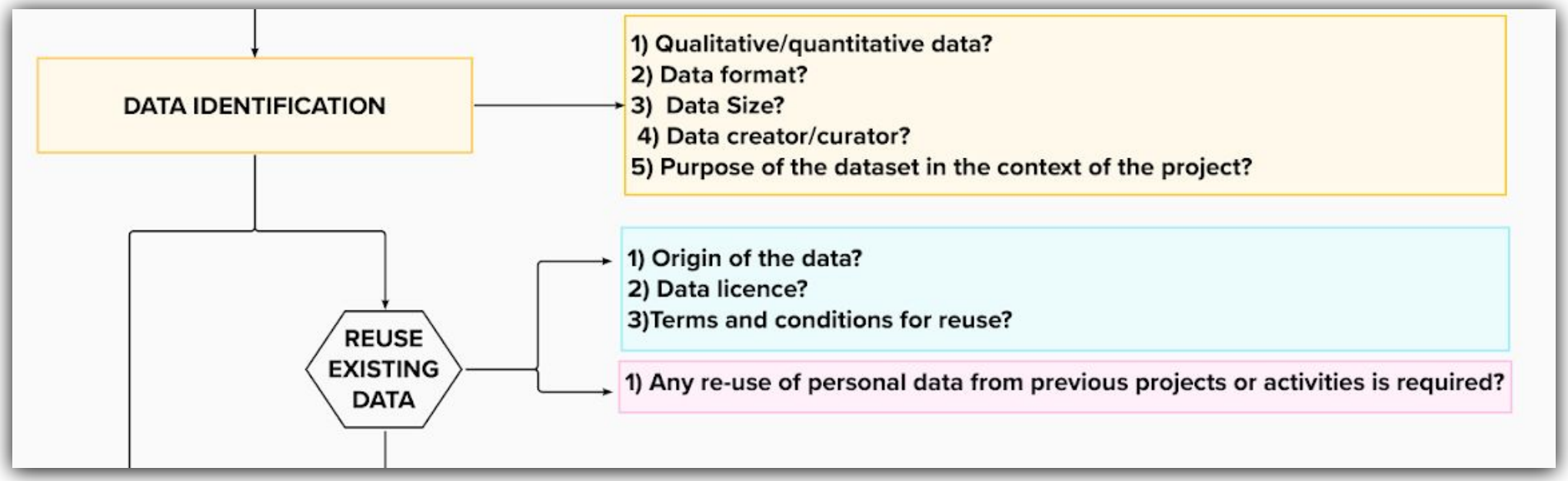
Reusable

Dati corredati da una buona documentazione in modo da poter essere interpretati correttamente, replicati e/o combinati anche in contesti diversi.

- Readme files e documentazione
- Fonte e contesto di provenienza dei dati
- Strumenti necessari per riprodurre i risultati
- Licenze d'uso

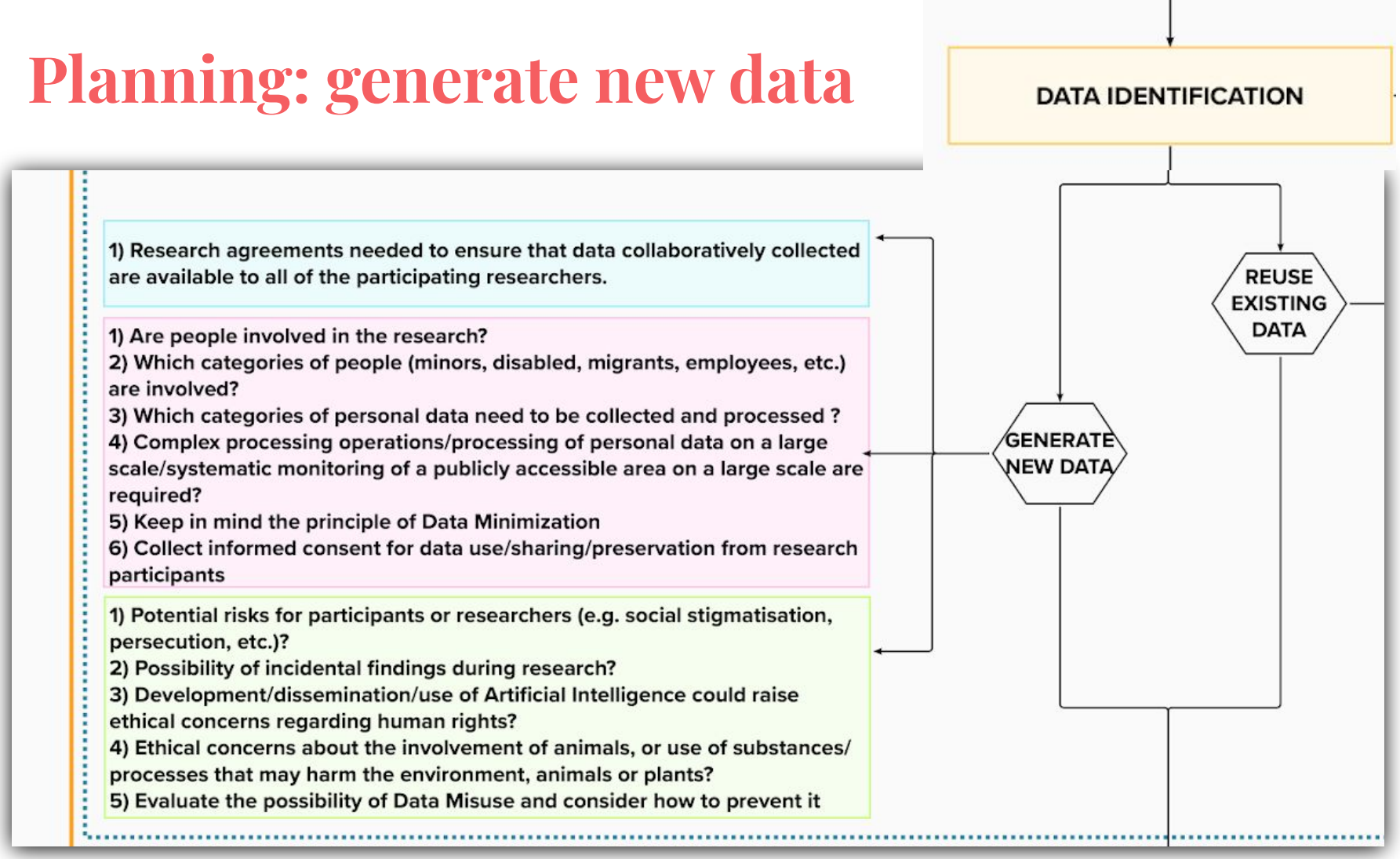
https://doi.org/10.5281/zenodo.8383693

Planning: reuse existing data



1) It applies when using data from other experiments/collaborations

Planning: generate new data



Privacy / Confidentiality:

- Experiments involving people (bio-physics, nuclear medicine)
- Personal data treatment

Ethics:

- Potential risk (from radiation sources, radiation activation)
- Data misuse