October 19 - 25, 2024

CHEP 2024

Conference on Computing in High Energy and Nuclear Physics

SO FAIR SO GOOD

the INFN strategy for Data Stewardship

Stefano Bianco, Daniele Bonacorsi, Concezio Bozzi, Luca Dell’Agnello, Luciano Gaido, Francesca Marchegiani, Irene Piergentili, **Lorenzo Rinaldi**, Stefano Dal Pra

Skills 4 eosc

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

INFN

CNAF, LNGS, LNF, Ferrara, Torino

Outline

- Introduction

  FAIR principles in High Energy and Nuclear Physics

- The INFN and the Data Stewards team

  People and connections

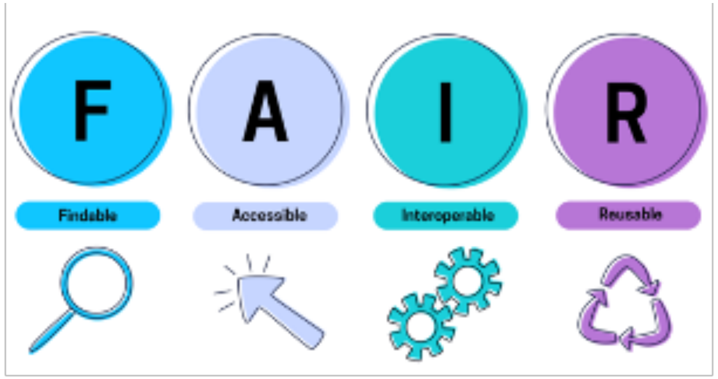
- The INFN Decision Tree for Data Management

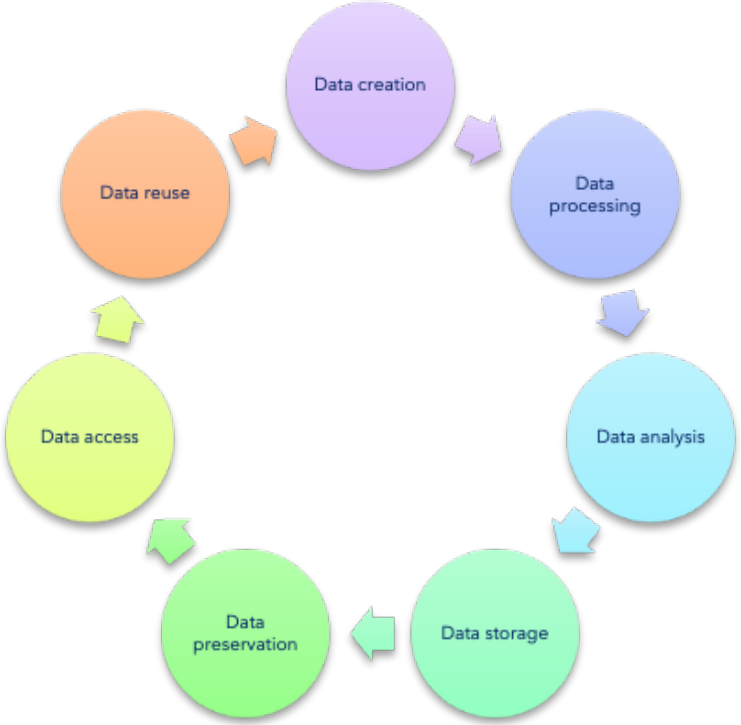
  How to draft a Data Management Plan

- 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 entities 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
- OpenAccess policies

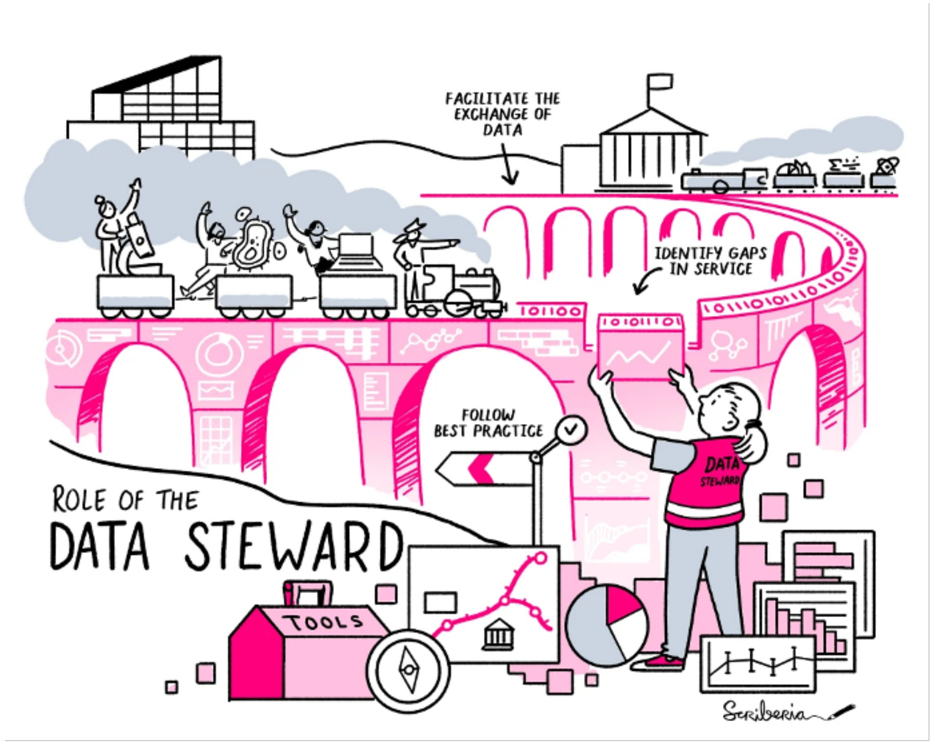
Many leading institutions for OpenScience (CERN, GSI, …)

What about small communities or individual experiments?
Who can help them?

The role of Data Steward

The Data Steward is a new professional figure

- 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

Characterization of data in HENP

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 entities

Data ownership

different funding entities 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

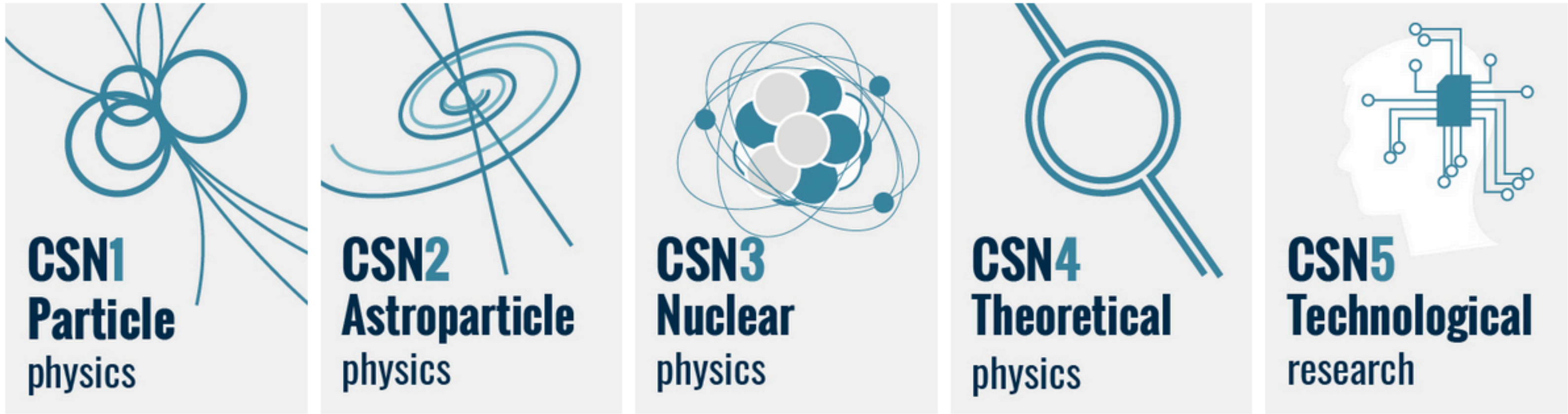
"Proprietary" software: acquisition, processing, and reading software developed within each scientific collaboration to meet their own purposes and needs

Duration of experiments: 5 – 20 years

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

INFN

Leading research activity in experimental and theoretical Physics:
- 5 research lines (National Scientific Committees)



CSN1 Particle physics

CSN2 Astroparticle physics

CSN3 Nuclear physics

CSN4 Theoretical physics

CSN5 Technological research

Many local divisions, national labs and computing centers

Participation in large international collaborations
AND
support small communities and single researchers

Why does INFN need a team of Data Stewards?

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 INFN Data Steward team

Stefano Dal Pra

Senior technologist @ INFN-CNAF

Master degree in Electronic Engineering

Francesca Marchegiani

Technologist @ INFN-LNGS

Master degree in Chemistry

Irene Piergentili

Fellowship for technological research @INFN-LNF

Master degree in Archive and Library theory and management

Lorenzo Rinaldi

Associate Professor @ UNIBO and affiliation @ INFN

PhD in Physics

XXX YYY

Multidisciplinarity, experience in many transversal fields

A newborn team, waiting for new members
Organization of the team is on the way

User support plan

The main objective is to support small research groups

- Data Management Plan drafting, according to a precise check-list
- publication (which Open Access level?)

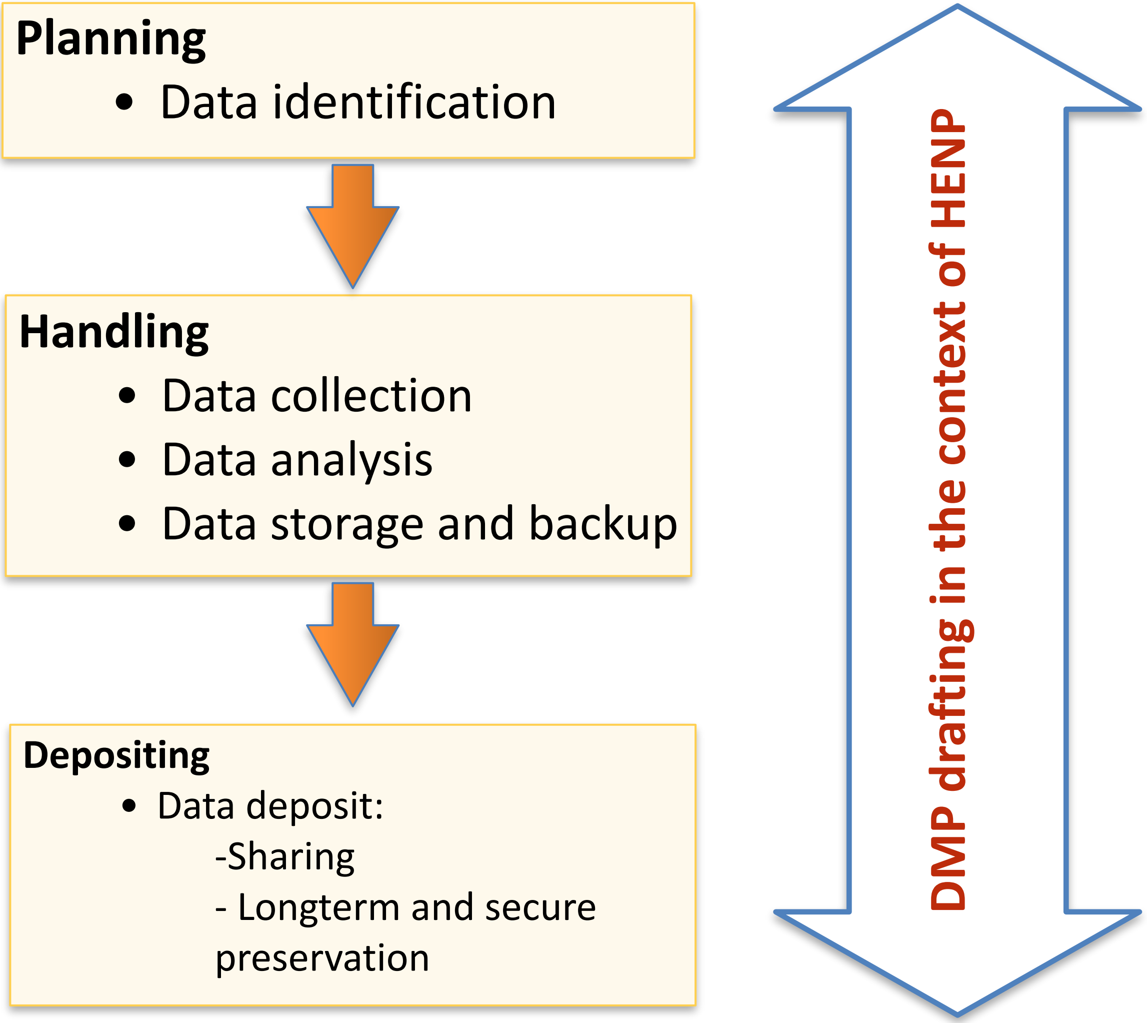
Target:

Researchers with no or few knowledge of FAIR principles

Different levels of support

Researchers (highly) familiar with FAIR principles

The Decision Tree for Data Management

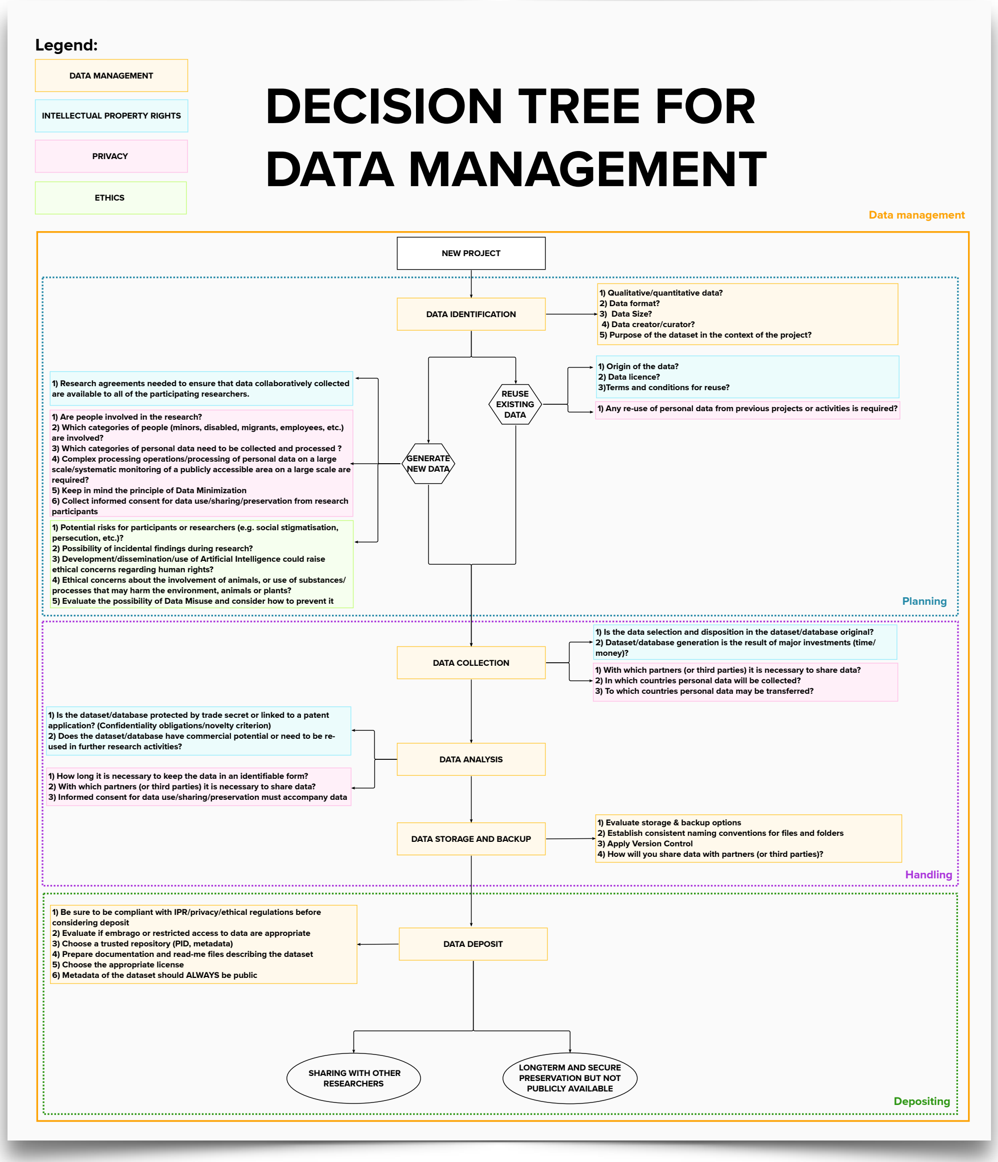


Planning
- Data identification

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

Depositing
- Data deposit:
  -Sharing
  - Longterm and secure preservation

DMP drafting in the context of HENP



1) Qualitative/quantitative data?
2) Data format?
3) Data Size?
4) Data creator/curator?
5) Purpose of the dataset in the context of the project?

1) Origin of the data?
2) Data licence?
3)Terms and conditions for reuse?

1) Any re-use of personal data from previous projects or activities is required?

1) Research agreements needed to ensure that data collaboratively collected are available to all of the participating researchers.

1) Are people involved in the research?
2) Which categories of people (minors, disabled, migrants, employees, etc.) are involved?
3) Which categories of personal data need to be collected and processed ?
4) Complex processing operations/processing of personal data on a large scale/systematic monitoring of a publicly accessible area on a large scale are required?
5) Keep in mind the principle of Data Minimization
6) Collect informed consent for data use/sharing/preservation from research participants

1) Potential risks for participants or researchers (e.g. social stigmatisation, persecution, etc.)?
2) Possibility of incidental findings during research?
3) Development/dissemination/use of Artificial Intelligence could raise ethical concerns regarding human rights?
4) Ethical concerns about the involvement of animals, or use of substances/processes that may harm the environment, animals or plants?
5) Evaluate the possibility of Data Misuse and consider how to prevent it

1) Is the data selection and disposition in the dataset/database original?
2) Dataset/database generation is the result of major investments (time/money)?

1) With which partners (or third parties) it is necessary to share data?
2) In which countries personal data will be collected?
3) To which countries personal data may be transferred?

1) Is the dataset/database protected by trade secret or linked to a patent application? (Confidentiality obligations/novelty criterion)
2) Does the dataset/database have commercial potential or need to be re-used in further research activities?

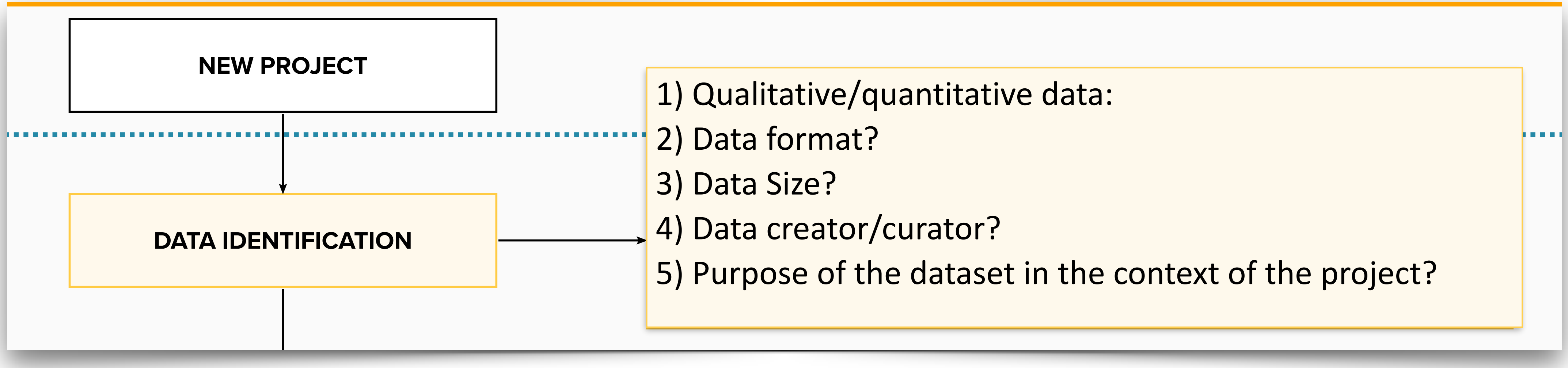
1) How long it is necessary to keep the data in an identifiable form?
2) With which partners (or third parties) it is necessary to share data?
3) Informed consent for data use/sharing/preservation must accompany data

1) Evaluate storage & backup options
2) Establish consistent naming conventions for files and folders
3) Apply Version Control
4) How will you share data with partners (or third parties)?

1) Be sure to be compliant with IPR/privacy/ethical regulations before considering deposit
2) Evaluate if embrago or restricted access to data are appropriate
3) Choose a trusted repository (PID, metadata)
4) Prepare documentation and read-me files describing the dataset
5) Choose the appropriate license
6) Metadata of the dataset should ALWAYS be public

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

Planning: data identification

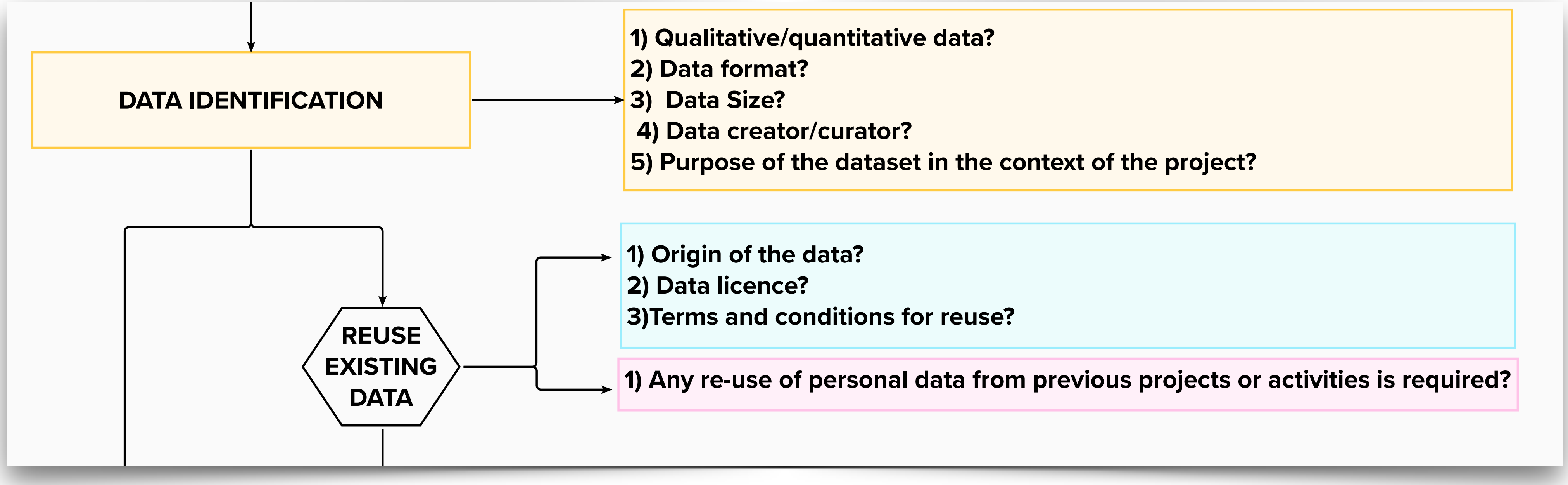


1) Qualitative/quantitative data:
2) Data format?
3) Data Size?
4) Data creator/curator?
5) Purpose of the dataset in the context of the project?

Some HENP examples in context:

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

Planning: reuse existing data



DATA IDENTIFICATION

1) Qualitative/quantitative data?
2) Data format?
3) Data Size?
 4) Data creator/curator?
5) Purpose of the dataset in the context of the project?

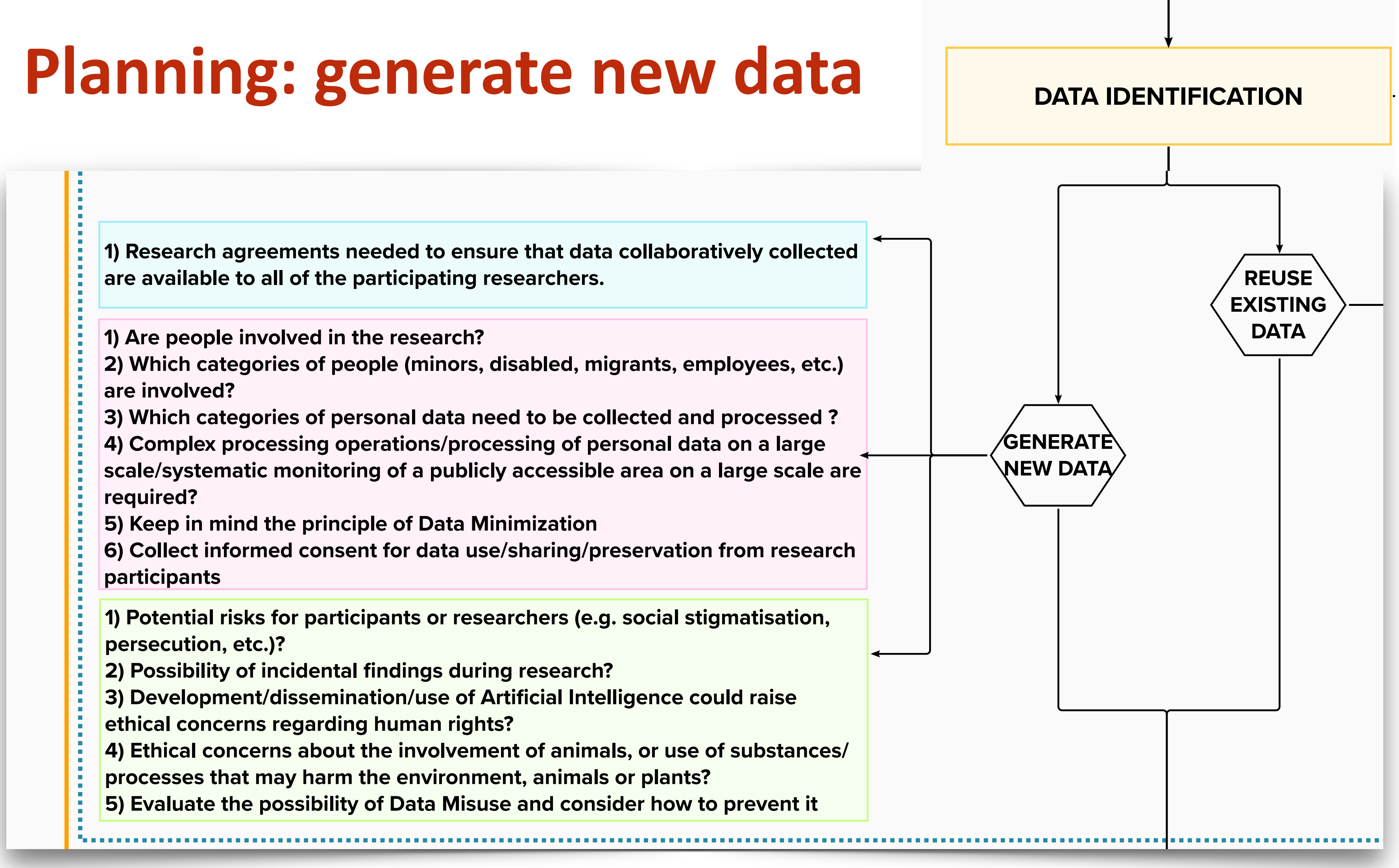
REUSE EXISTING DATA

1) Origin of the data?
2) Data licence?
3)Terms and conditions for reuse?

1) Any re-use of personal data from previous projects or activities is required?

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

Planning: generate new data



1) Research agreements needed to ensure that data collaboratively collected are available to all of the participating researchers.

1) Are people involved in the research?
2) Which categories of people (minors, disabled, migrants, employees, etc.) are involved?
3) Which categories of personal data need to be collected and processed ?
4) Complex processing operations/processing of personal data on a large scale/systematic monitoring of a publicly accessible area on a large scale are required?
5) Keep in mind the principle of Data Minimization
6) Collect informed consent for data use/sharing/preservation from research participants

1) Potential risks for participants or researchers (e.g. social stigmatisation, persecution, etc.)?
2) Possibility of incidental findings during research?
3) Development/dissemination/use of Artificial Intelligence could raise ethical concerns regarding human rights?
4) Ethical concerns about the involvement of animals, or use of substances/ processes that may harm the environment, animals or plants?
5) Evaluate the possibility of Data Misuse and consider how to prevent it

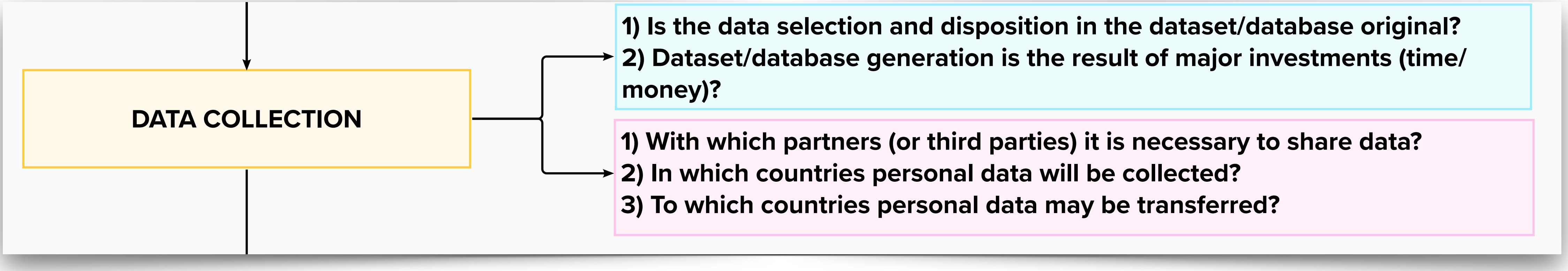
Privacy / Confidentiality:

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

Ethics:

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

Handling: data collection



DATA COLLECTION

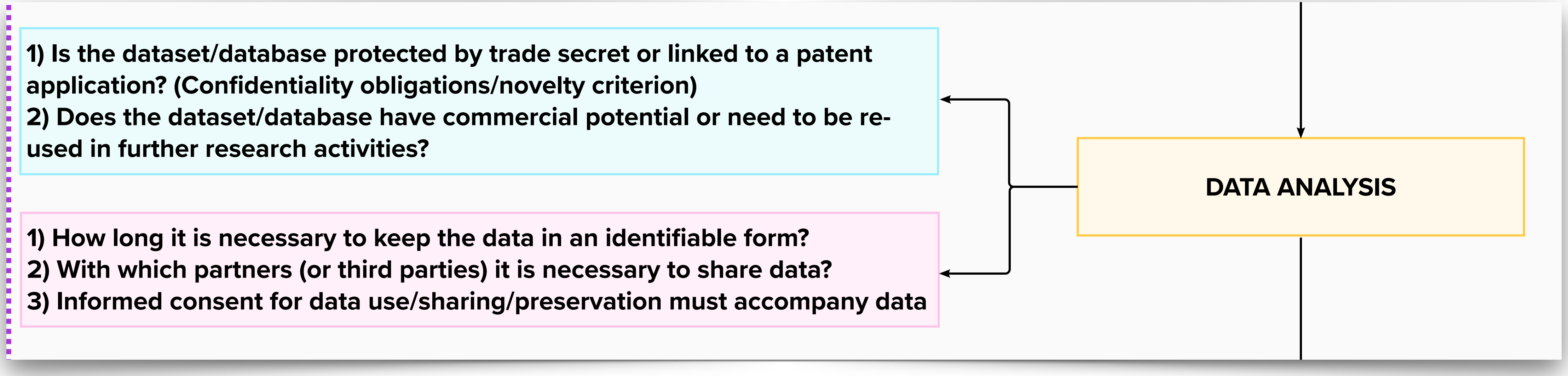
1) Is the data selection and disposition in the dataset/database original?
2) Dataset/database generation is the result of major investments (time/money)?

1) With which partners (or third parties) it is necessary to share data?
2) In which countries personal data will be collected?
3) To which countries personal data may be transferred?

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…

1) Collaboration with private companies, Technology Transfer
2) International collaboration (same rules?)—> research agreement

Handling: data analysis



DATA ANALYSIS

1) Is the dataset/database protected by trade secret or linked to a patent application? (Confidentiality obligations/novelty criterion)
2) Does the dataset/database have commercial potential or need to be re-used in further research activities?

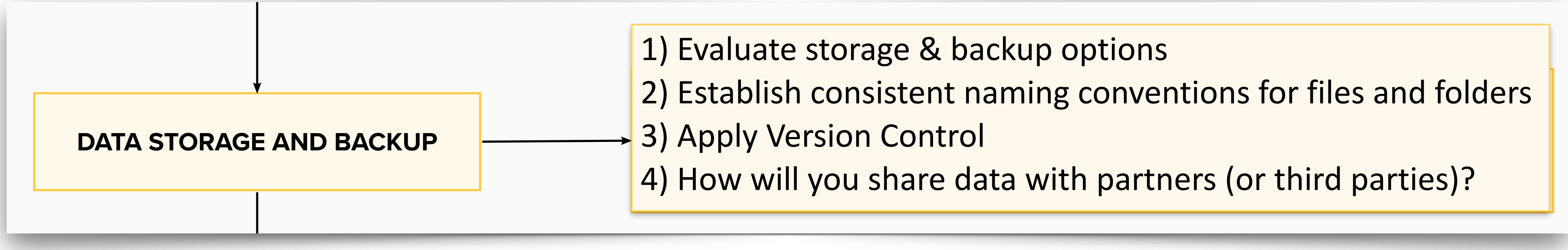
1) How long it is necessary to keep the data in an identifiable form?
2) With which partners (or third parties) it is necessary to share data?
3) Informed consent for data use/sharing/preservation must accompany data

Public research or collaboration with private companies, Technology Transfer ?

Privacy/Confidentiality:

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

Handling: data storage and backup


DATA STORAGE AND BACKUP

1) Evaluate storage & backup options
2) Establish consistent naming conventions for files and folders
3) Apply Version Control
4) How will you share data with partners (or third parties)?

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…)

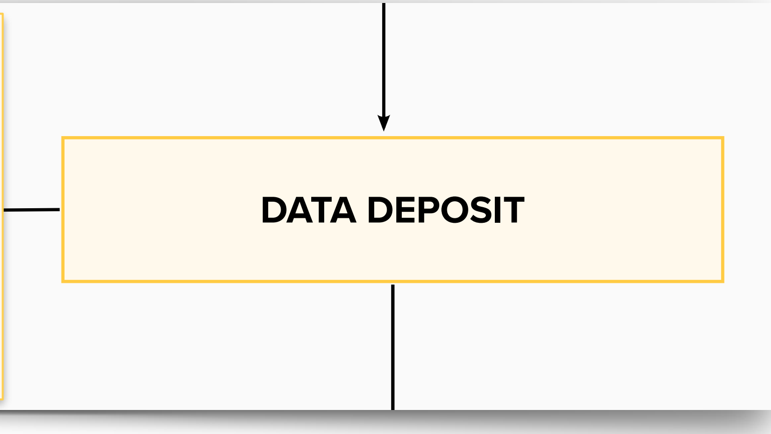
3) For the Software: Git-Github, Gitlab, **Baltig***

4) remote&secure access (IAM or other secure authentication)
   Public storage endpoint via WEB interface (for OpenData)

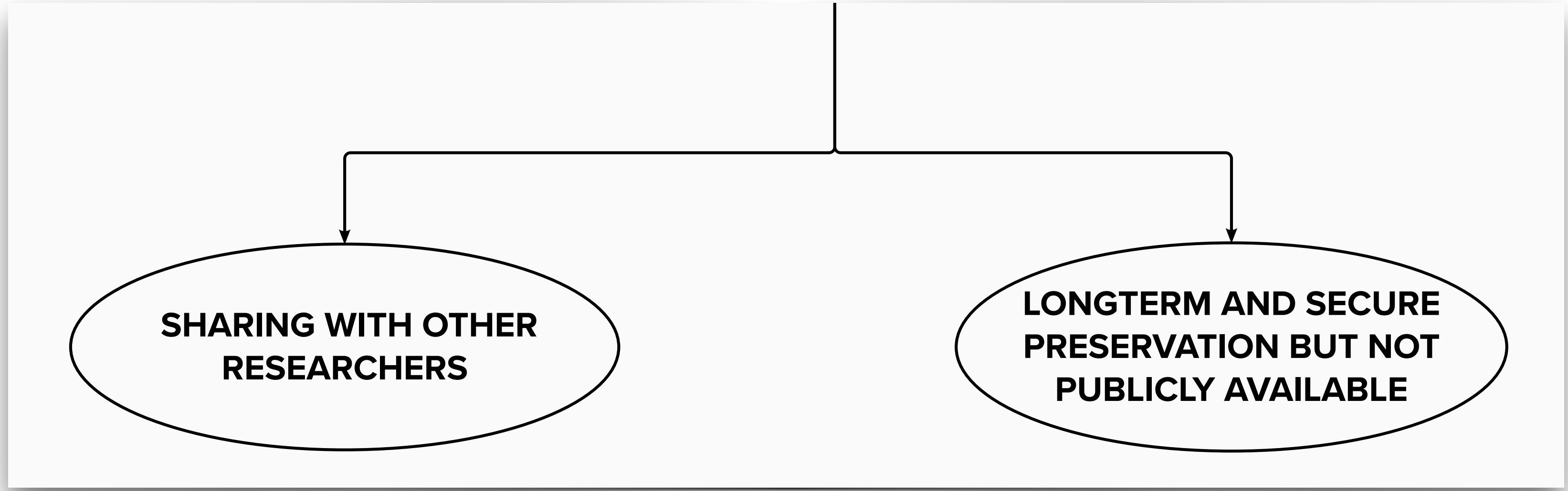
*** managed by INFN**

Depositing

1) Be sure to be compliant with IPR/privacy/ethical regulations before considering deposit
2) Evaluate if embargo or restricted access to data are appropriate
3) Choose a trusted repository (PID, metadata)
4) Prepare documentation and read-me files describing the dataset
5) Choose the appropriate license
6) Metadata of the dataset should ALWAYS be public



INFN Open Access Repository:
https://www.openaccessrepository.it/


SHARING WITH OTHER RESEARCHERS
LONGTERM AND SECURE PRESERVATION BUT NOT PUBLICLY AVAILABLE

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 has now a team of Data Stewards

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

Next: Get in touch with real use cases to implement and improve the support checklist

THANK YOU!

Lorenzo Rinaldi

Bologna University & INFN

lorenzo.rinaldi@unibo.it

www.unibo.it

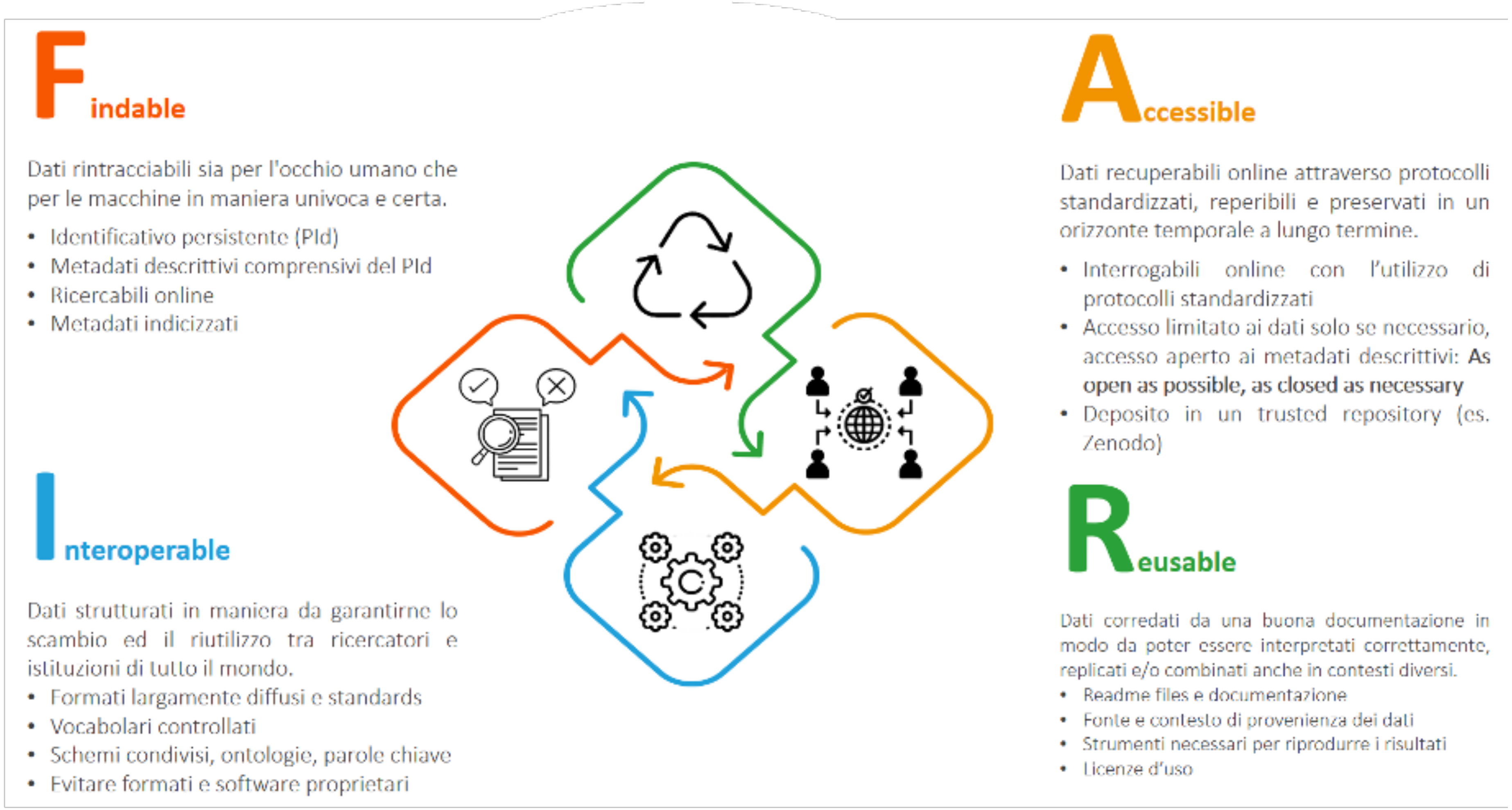
Skills 4 eosc

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

INFN

Backup

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

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

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. Zenodo)

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