Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules

Data life cycle and Archiving

Yonny CARDENAS

cardenas@cc.in2p3.fr

24 September 2021

Scientific data life cycle: motivations

CC-IN2P3:

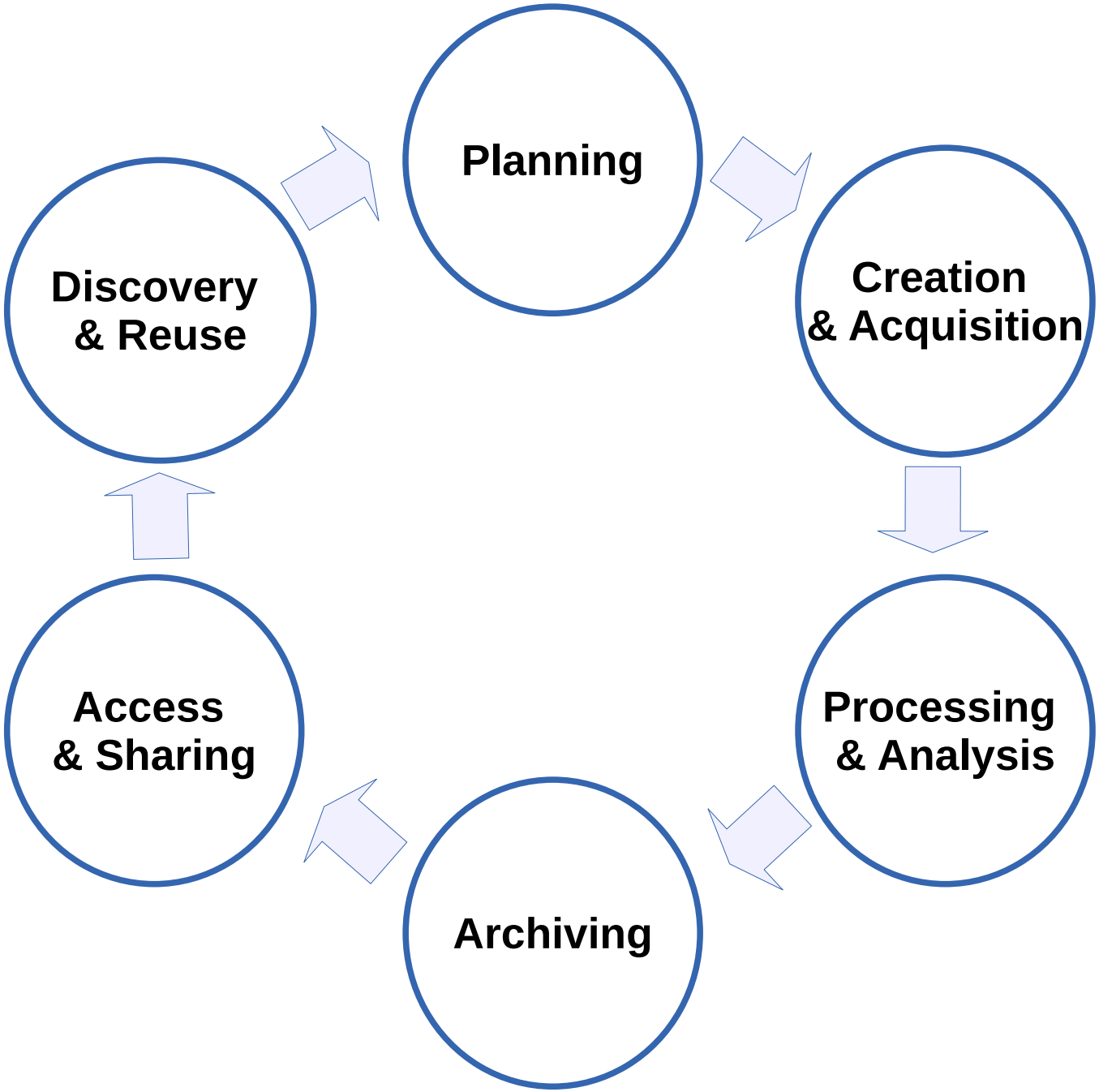
- *What will be the future of the data?* réunion des expériences 2015
  https://indico.in2p3.fr/event/10875
- a project become inactive, how about the data?
  - data become orphan and/or obsolete over time
  - nothing can be done !

- question important for the «current» and «future» projects
- data life cycle improvement
  - Archival service (OAIS) to ensure data reusability
  - Data Management Plan

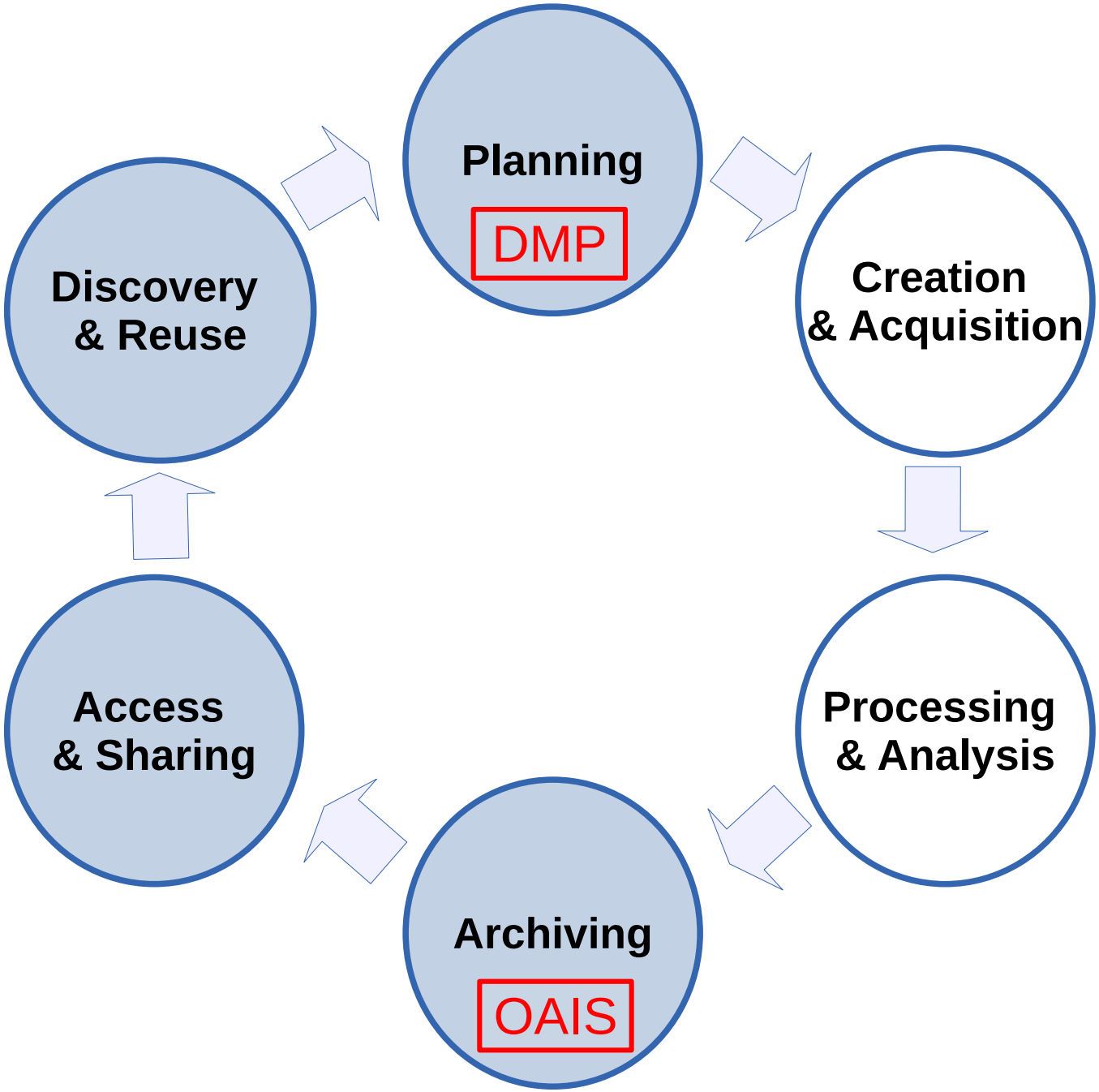
Global context:

- movement Open Data and Open Science
  - scientific research data must be accessible to all
- funding agencies constrains access to resources
  - principles F.A.I.R (findability, accessibility, interoperability, and reusability)
  - Data Management Plan

Scientific data live cycle



Scientific data live cycle



Data Management Plan

A DMP describes the data management life cycle for the data.

A DMP should include information on:

- general description (purpose of the data)
- the handling data during & after the end of the project (responsibility)
- data usage scenarios
- making data FAIR (including metadata)
- allocation of resources (costs)
- legal and ethical aspects
- how data will be curated & preserved (including after the end of the project)

Data Management Plan at IN2P3

- It is not an administrative form to fill !
- Keep in mind that is a planning process (make decisions)
- not all questions are relevant to your particular case
  - unanswered questions are possible
- Strategy proposed by astrophysics community
  - Leibniz Institute for Astrophysics Potsdam
  - generic questionnaire as guide (similar to an interview)
  - no references to specific technologies or services
  - the main objective is to encourage a global reflection
    - different aspects of data management

- Full questionnaire (printer version) at https://irods.in2p3.fr/dmp
- On line versions: https://dmp.in2p3.fr or https://dmp.opidor.fr

IN2P3 DMP Questionnaire

General

→ Topic
- Research field
- Project schedule
- Project coordination
- Project partners
- Funding
- Other requirements I
- Other requirements II

Content classification

→ Datasets
- Data origin
- Reuse
- Reproducibility

Technical classification

→ Date collection
- Data size
- Formats
- Tools
- Versioning

Data usage

→ Usage scenarios
- Data organisation
- Data storage and security
- Interoperability
- Data sharing and re-use
- Collaborative work
- Quality assurance
- Data integration
- Costs

Metadata and referencing

→ Metadata
- Metadata costs
- Structure, granularity, and referencing
- Persistent Identifiers (PIDs)
- PIDs costs

Storage and long-term preservation

→ Selection
- Long-term preservation
- Long-term preservation costs

Legal and ethics

→ General legal issues
- Personal data
- Data protection
- Sensitive data
- Other sensitive data
- Sensitive data costs
- Official approval
- Intellectual property rights I
- Intellectual property rights II
- Intellectual property rights costs

Archiving Service

Service for the management and preservation of large amount of scientific data.

Objectives:

- Data life cycle improvement
- Preserve scientific data to ensure that remains accessible and reusable.
- Respond to the specific requirements of IN2P3 experiments
- Improve CC-IN2P3 storage services offer

Backup vs Archiving

Backup	Archiving
• operational continuity	• patrimony preservation
• data in production • data modifiable, in progress • all data is potentially concerned • frequency:many times(versions)	• precious or finished data • frozen, validated data • only selected data • frequency:one time
• short-term content retention (hours,days,weeks,months)	• long-term content retention (years,decades,…)
• use proprietary technologies • strong dependencies • low interoperability • create and restitution time must be (very)short	• use open and free technologies • weak dependencies • full interoperability • create and restitution time are not critical
• automatic process (humanless) • automatic data removing	• semi-automatic (curation) • only manual data removing
• Internal (operational)	• external: dissemination

Digital Archiving: some principles

- predict the scene of a great disaster
  - recover information directly from storage support (e.g. tape)
- information packages
  - wrapped: data object + metadata + administrative information
  - complain with standard specifications
  - self-contained and self-described
  - human-readable and machine-actionable
  - implementation: tar file
- strong reduction of technical dependencies
  - the minimum possible
  - use standard, open and widely known technologies
  - archive cannot depend of archive management software (disposable)
  - proprietary solutions are forbidden
- several copies
  - minimal two, three recommended
  - on different technologies (e.g. tape and disk)
  - on different locations (a copy more than 300 km away)

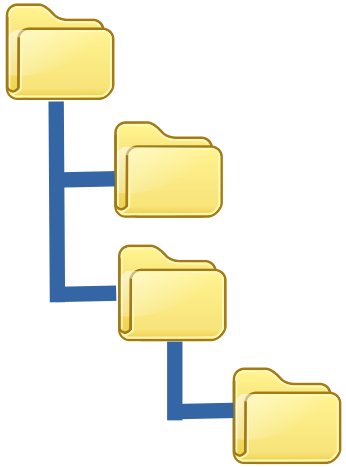
Archiving Service: features at IN2P3

- target large datasets (from several terabytes)
- based on the OAIS reference model
- implements European specifications E-ARK/CSIP
- compatible with F.A.I.R. process (producer: research project)
- preservation during a defined time period
- not addressed to administrative documents (not probatory value)
- on demand and adapted to the producer (scientific experiments)
- simplified procedures to cover all use cases
- uses the existing CC-IN2P3 infrastructure

Archiving research data

I want archive and disseminate my data !

Archiving research data

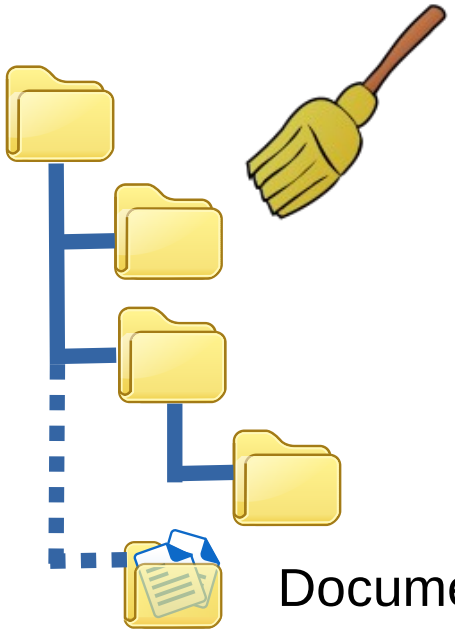


DATA

Metadata

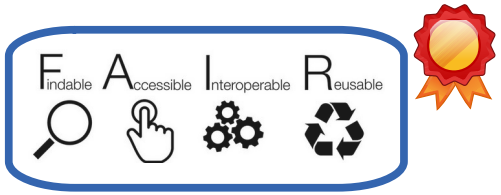
DMP

Cycle live data and Archiving



DATA

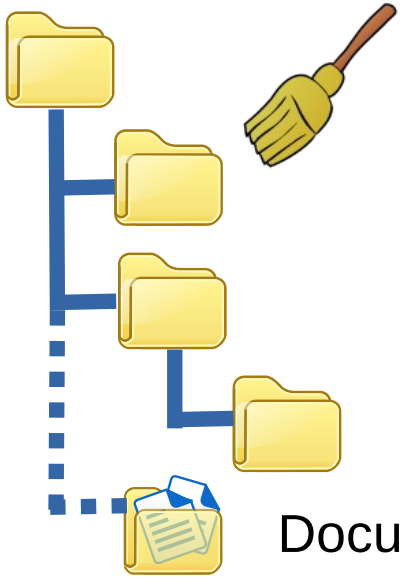
Documentation



Metadata

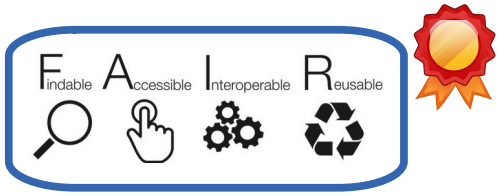
DMP

Cycle live data and Archiving



DATA

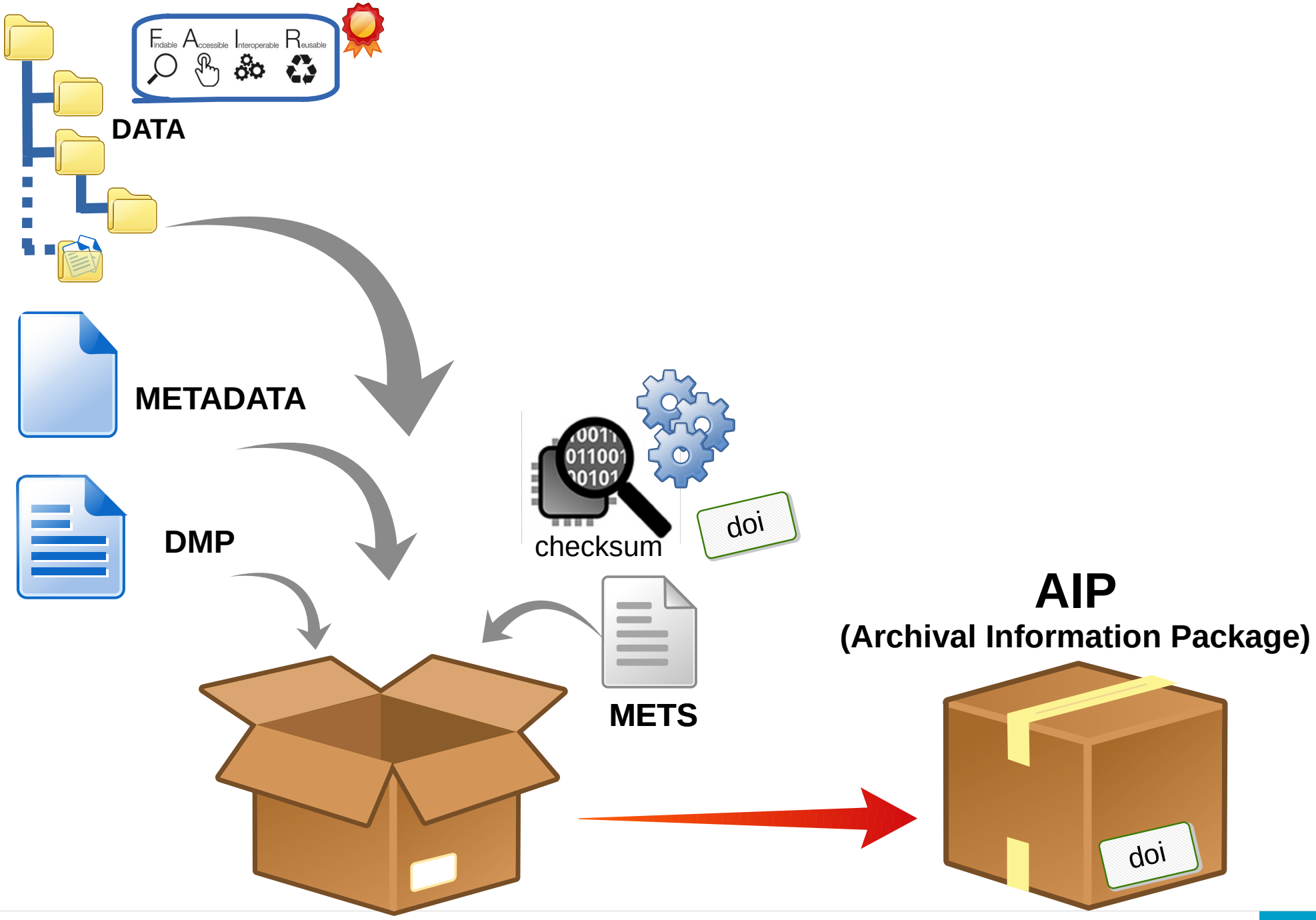
Documentation



Metadata

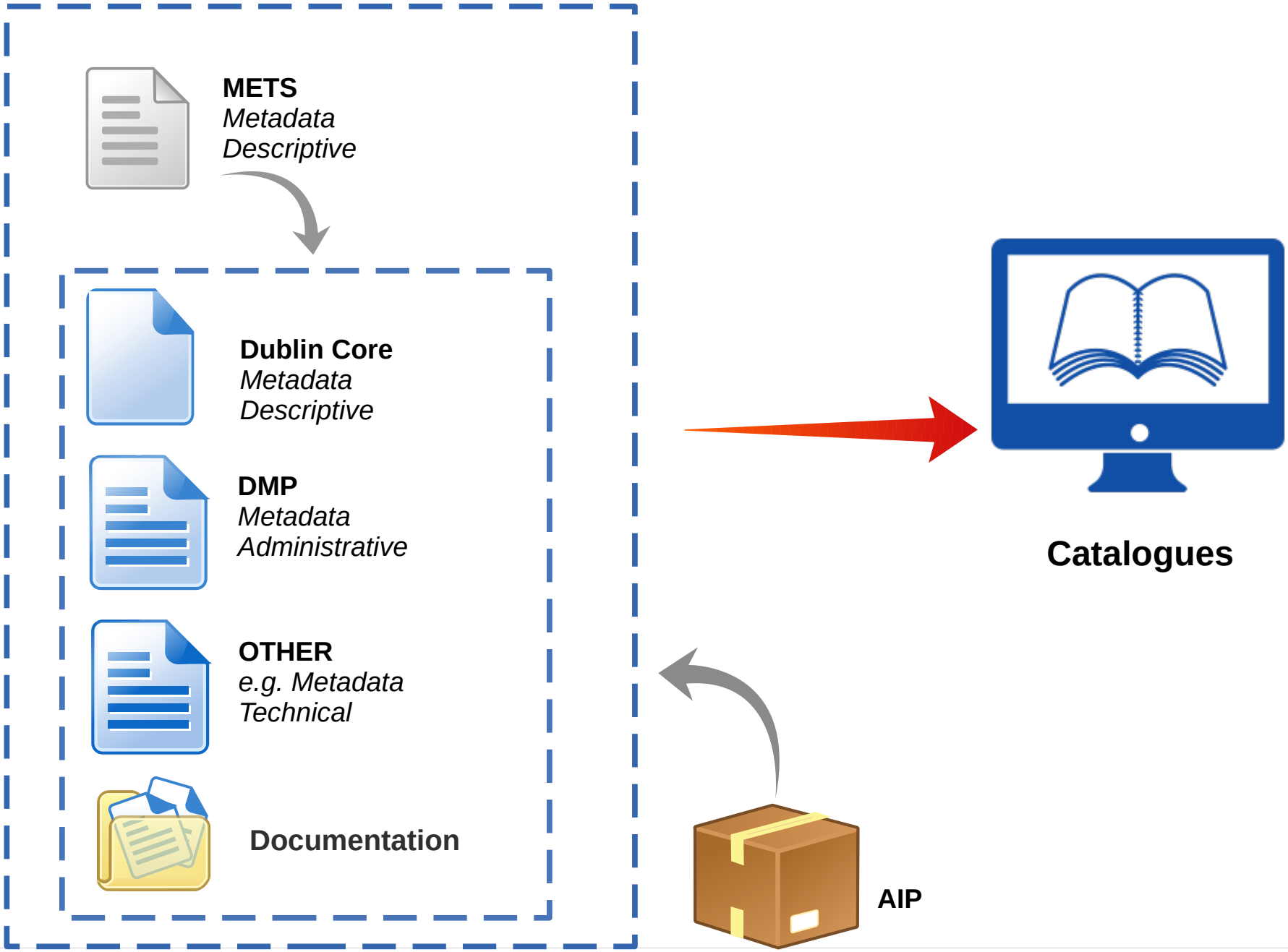
DMP

Archiving process



Archiving = preserve + dissemination

Using metadata and DMP



Archiving = preserve + dissemination

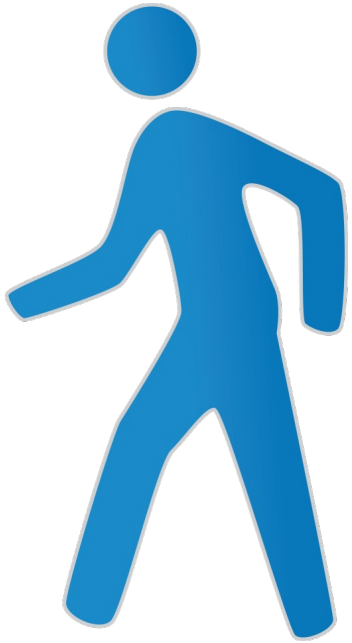
Catalogue

Archiving = preserve + dissemination

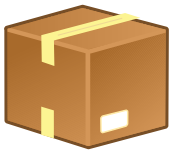
Archive

Catalogue

Search



User



DIP

(Dissemination Information Package)

Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules

Data life cycle and Archiving

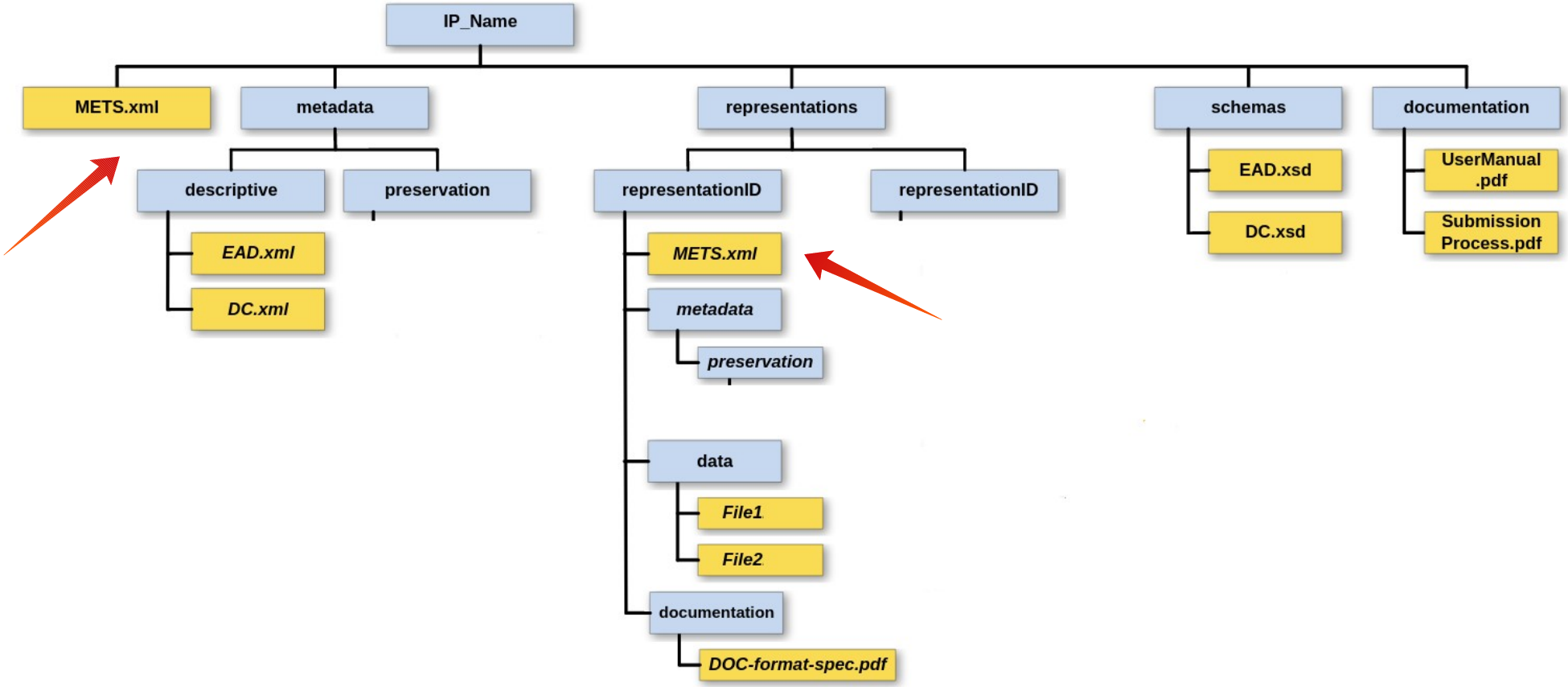
Yonny CARDENAS

cardenas@cc.in2p3.fr

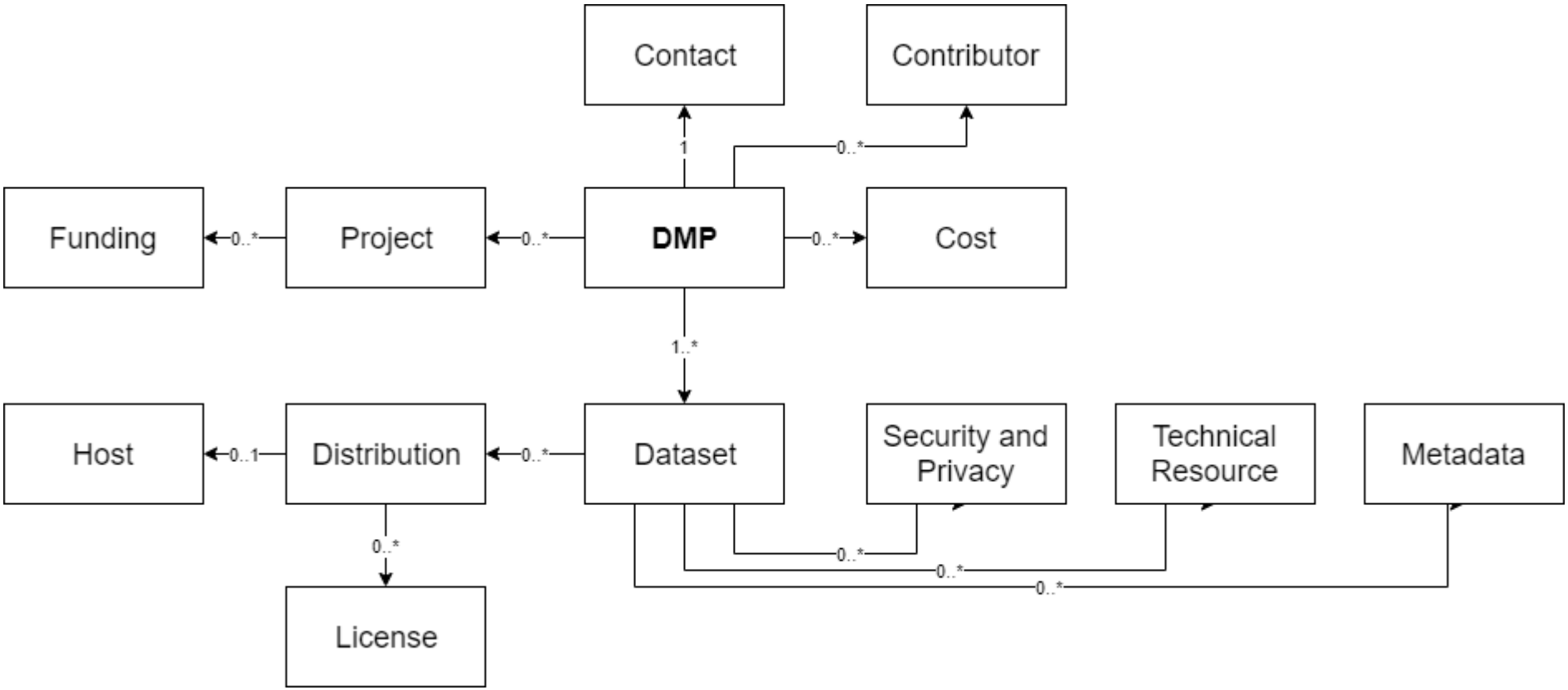
24 September 2021

Common Specification for Information Packages

CISP information package structure



DMP Common Standard Model



Archivage : planification

ICPSR

Inter-university Consortium for Political and Social Research

- https://www.icpsr.umich.edu
- DMP + letter of commitment
- *"A letter of commitment from ICPSR confirming that it will archive the data should accompany the plan."*