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

Using OAIS reference model for storing and preserving large amounts of scientific data

Yonny CARDENAS
cardenas@cc.in2p3.fr

Journées Calcul Données

15 December 2021

Motivations

- Research institutions or projects can produce large amounts of digital data
  - size dataset from some terabytes to petabytes
- Precious dataset
  - non-reproducible data
  - reproducible with disproportionate cost or effort
- When a project become inactive, how about the data?
  - data become orphan and/or obsolete over time 😒
  - nothing can be done !
  - What will be the future of the data?
- Principles F.A.I.R.
  - not specifically cite preservation as a requirement
  - management of data cycle life
  - early planification: DMP
- Key question for the «current» and «future» projects

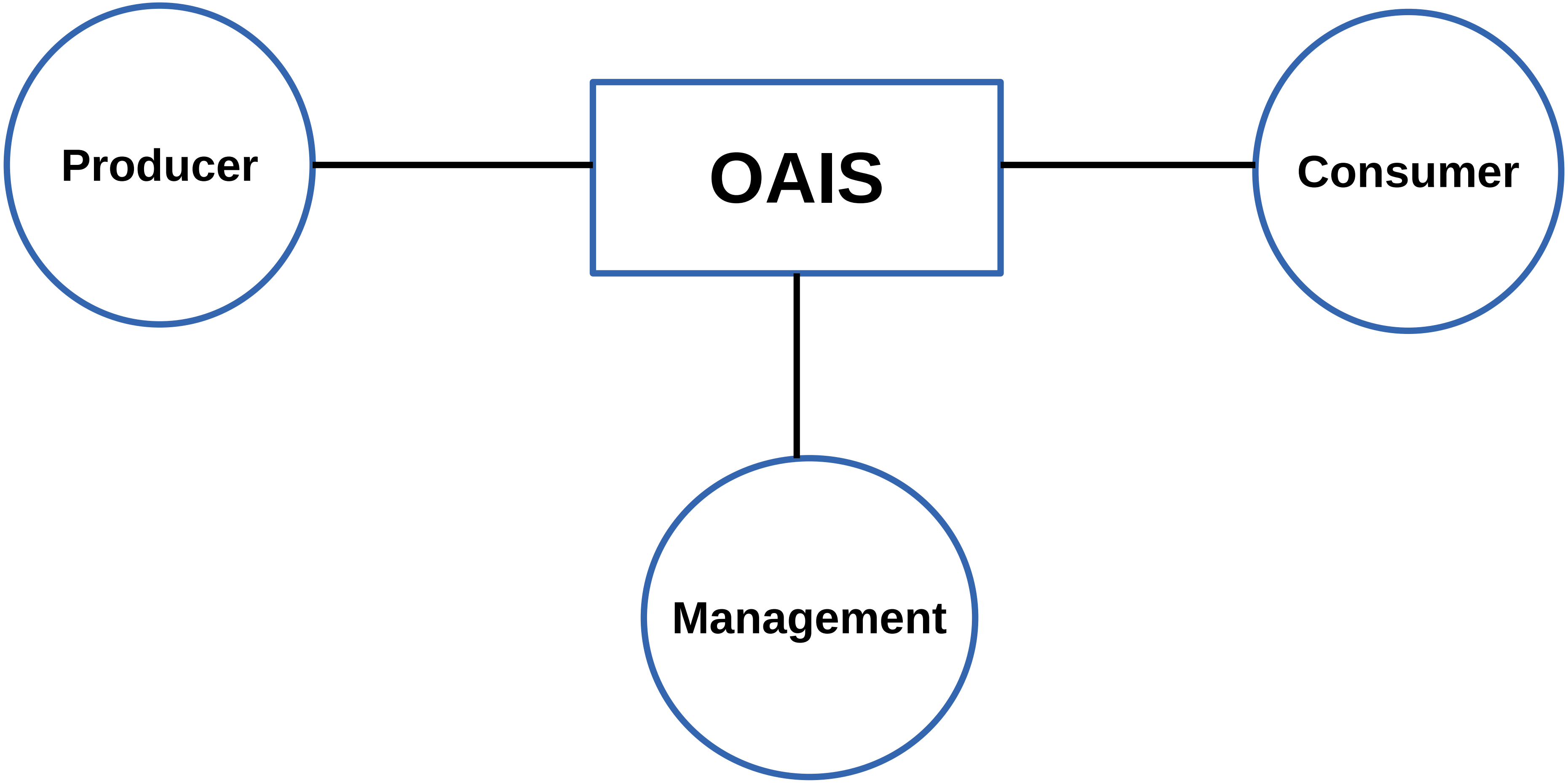
Backup vs Preservation

Backup	Preservation
• operational continuity	• patrimony
• 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

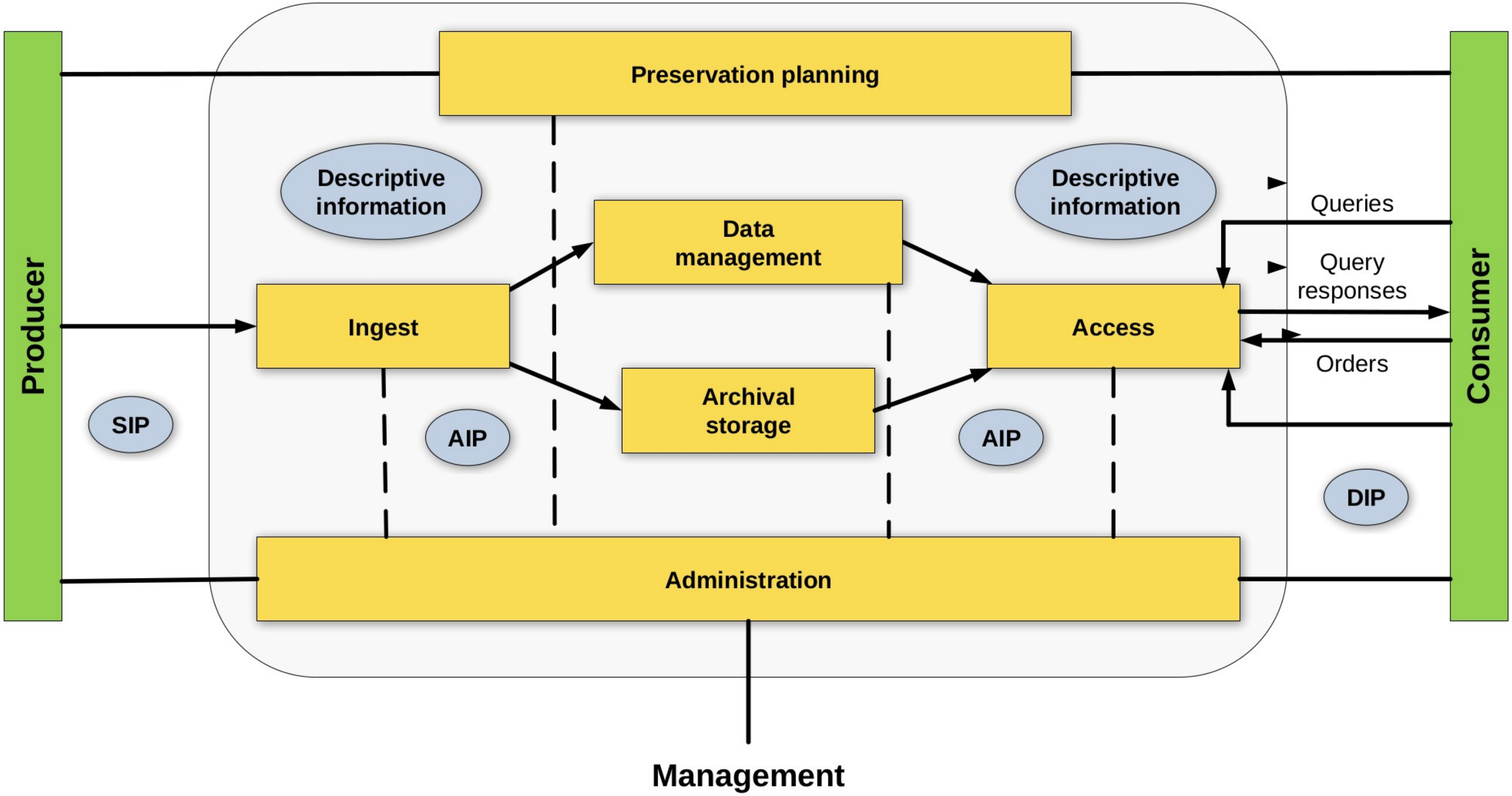
Digital Preservation: some principles

- 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
- 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)

Preservation: OAIS reference model

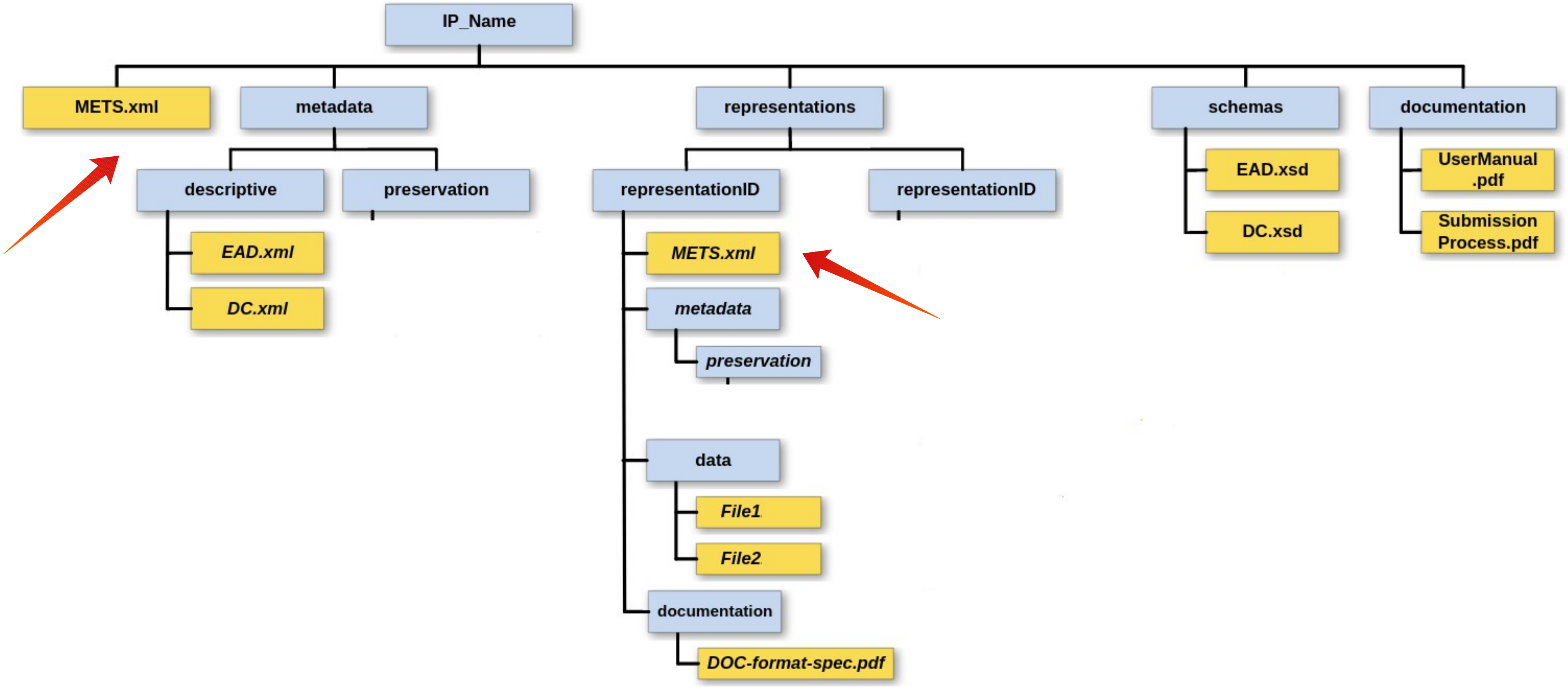


Preservation: OAIS reference model



Common Specification for Information Packages

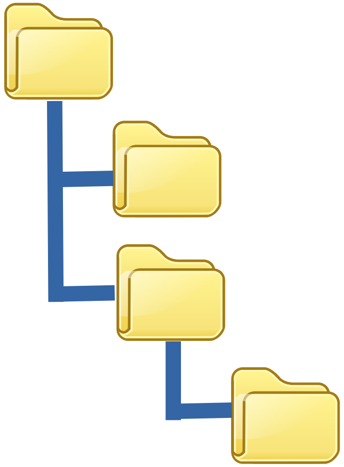
CISP information package structure



Preserving research data

I want preserve and disseminate my data !

Preserving research data

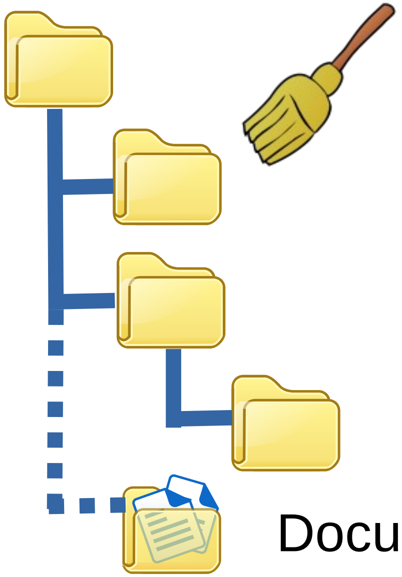


DATA

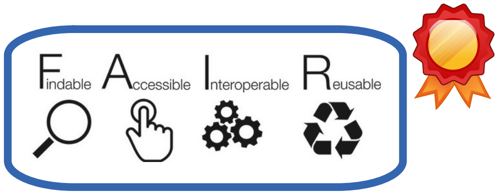
Metadata

DMP

Cycle live data and Preservation



DATA

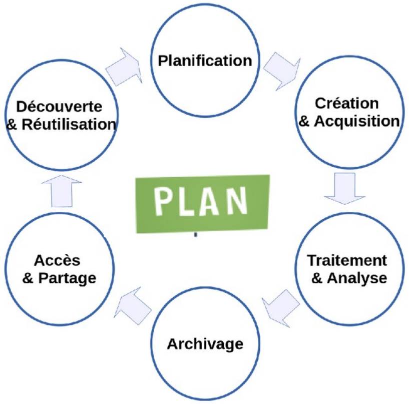


Documentation

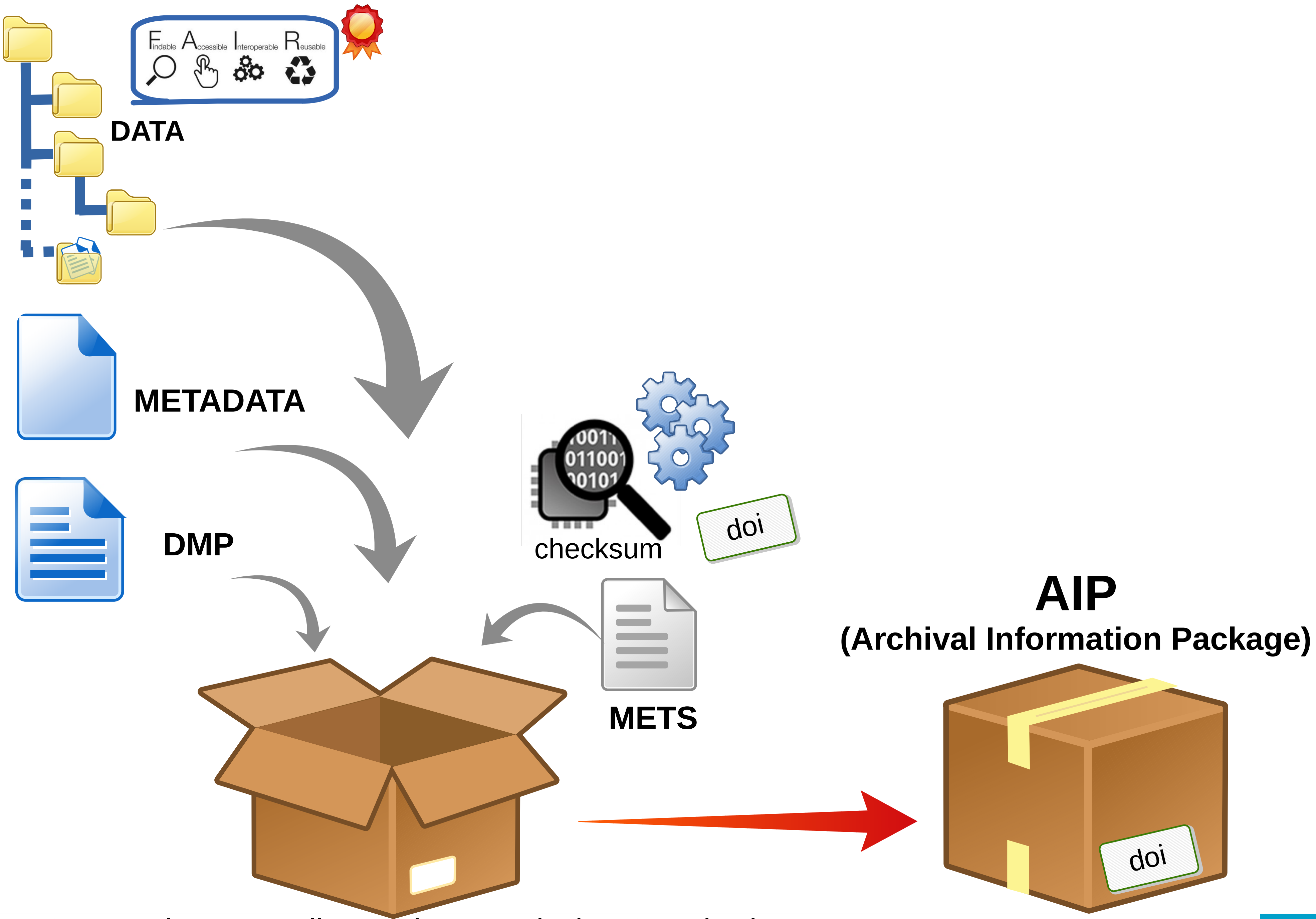
Metadata

title
format
relation
publisher
coverage
type
description
rights
language
source
contributor
creator
identifier
subject
date

DMP



Preserving process



DATA

METADATA

DMP

checksum

doi

METS

AIP

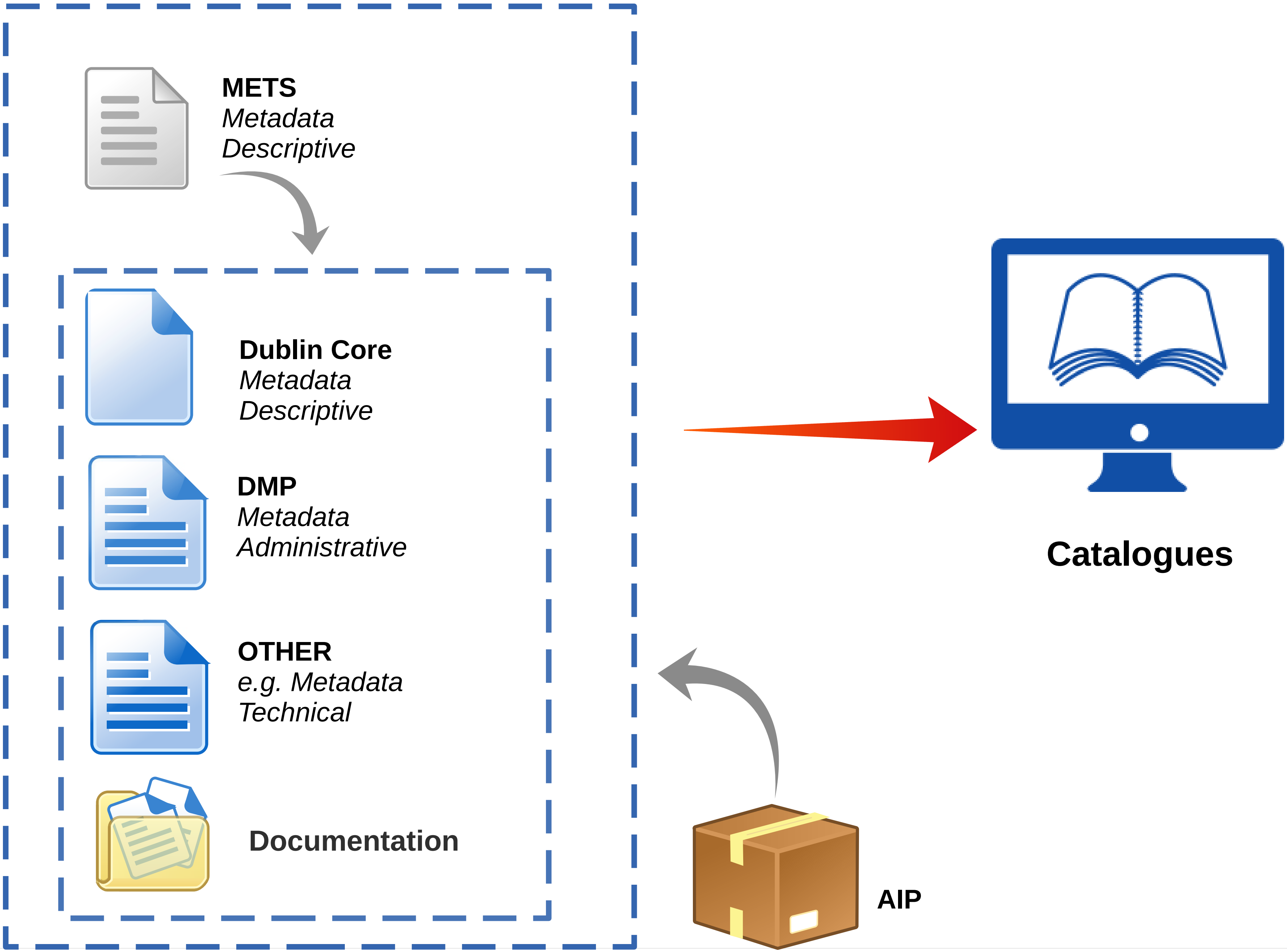
(Archival Information Package)

doi

METS: Metadata Encoding and Transmission Standard

Preservation + Dissemination

Using metadata and DMP



METS
Metadata
Descriptive

Dublin Core
Metadata
Descriptive

DMP
Metadata
Administrative

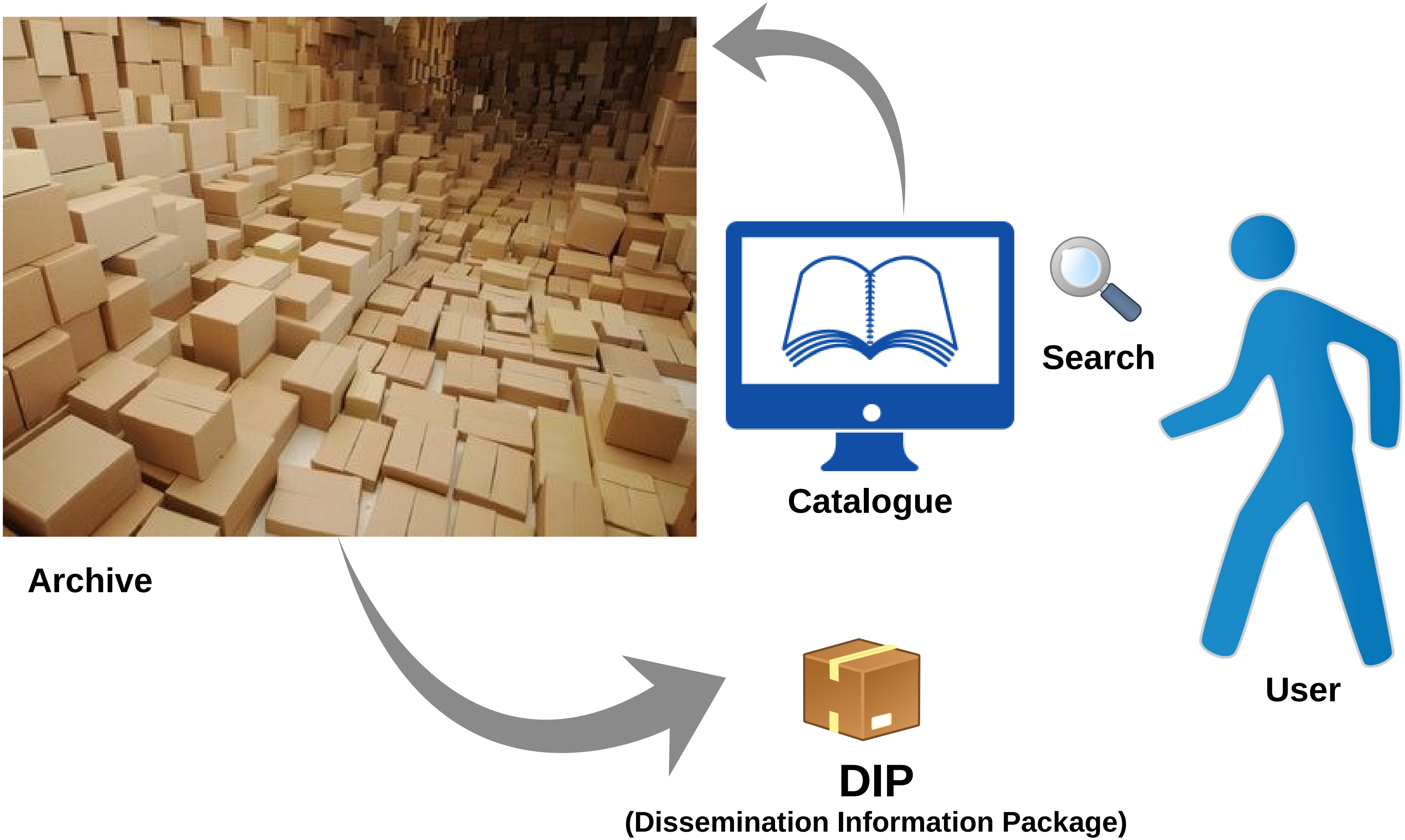
OTHER
e.g. Metadata
Technical

Documentation

Catalogues

AIP

Preservation + dissemination



Archive

Catalogue

Search

User

DIP

(Dissemination Information Package)

Summary

- target large datasets (from several terabytes to petabytes)
- light and flexible implementation of OAIS reference model
  - data curation based on F.A.I.R. process (producer: research project)
  - not addressed to administrative documents (not probatory value)
  - simplified procedures: e.g. ingest (not SEDA protocol)
  - accept all data formats (not migration)
- compliant with european specifications E-ARK/CSIP
  - interoperability
- implemented at CC-IN2P3 using existing infrastructure
  - IRODS
  - Tape library

Thanks to

Céline Guyon
Présidente de l'Association des Archivistes Français (AAF)

Laurent Duplouy
Chef du service multimédias du département de l’audiovisuel
Bibliothèque Nationale de France (BNF)

Lorène Béchard
Responsable fonctionnelle du système d'archivage électronique
CINES

Thomas Leibovici and **Patrice Lucas**
Phobos: Parallel Heterogeneous Object Store
CEA

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

Using OAIS reference model for storing and preserving large amounts of scientific data

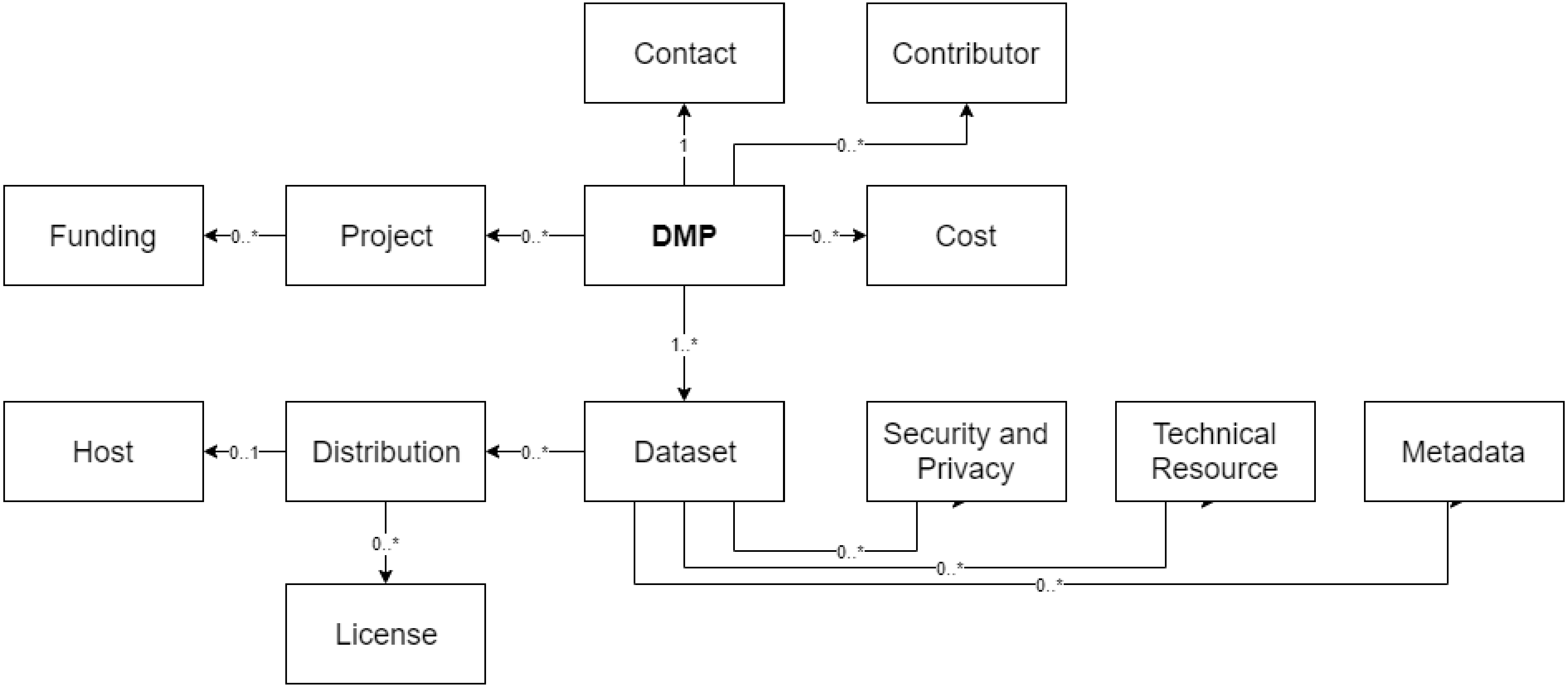
Yonny CARDENAS
cardenas@cc.in2p3.fr

Journées Calcul Données

15 December 2021

DMP Common Standard Model

RESEARCH DATA ALLIANCE



Contact

Contributor

1

0..*

Funding

0..*

Project

0..*

DMP

0..*

Cost

1..*

Host

0..1

Distribution

0..*

Dataset

Security and Privacy

Technical Resource

Metadata

0..*

0..*

0..*

0..*

License