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

Data Future – Les données du présent pour le futur

Yonny CARDENAS

cardenas at cc.in2p3.fr

RIP Data: quelle sélection, conservation et suppression des données de recherche?

1-2 octobre 2024, Aix-en-Provence

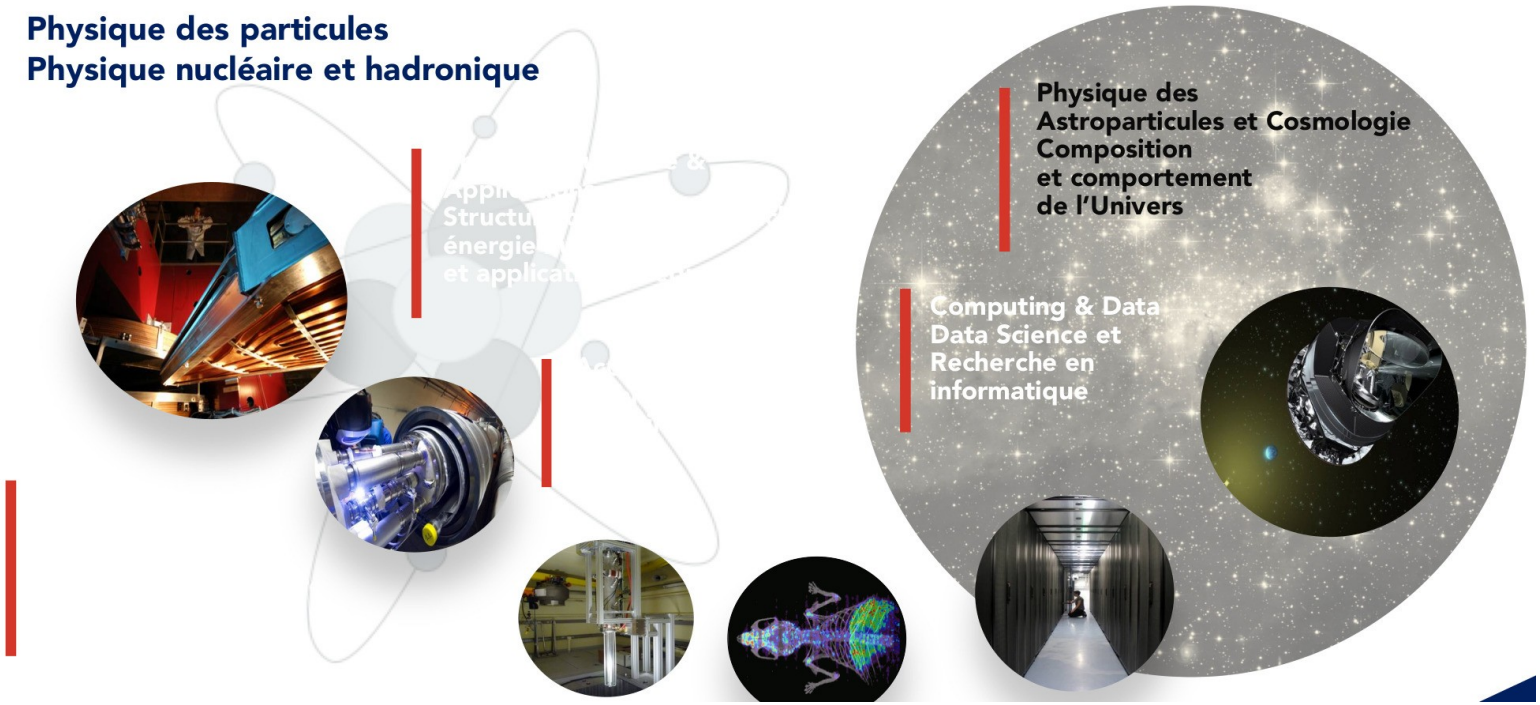
Le centre de calcul de l'IN2P3

Depuis 1986 au campus de la Doua à Lyon

Mission principale: déployer l'**infrastructure** et les **services informatiques** nécessaires aux chercheurs de l'IN2P3/CNRS

Services: **stockage** et traitement (calcul) de données scientifiques

Pas de service de préservation ou archivage des données !



Infrastructure: centre de calcul de l'IN2P3

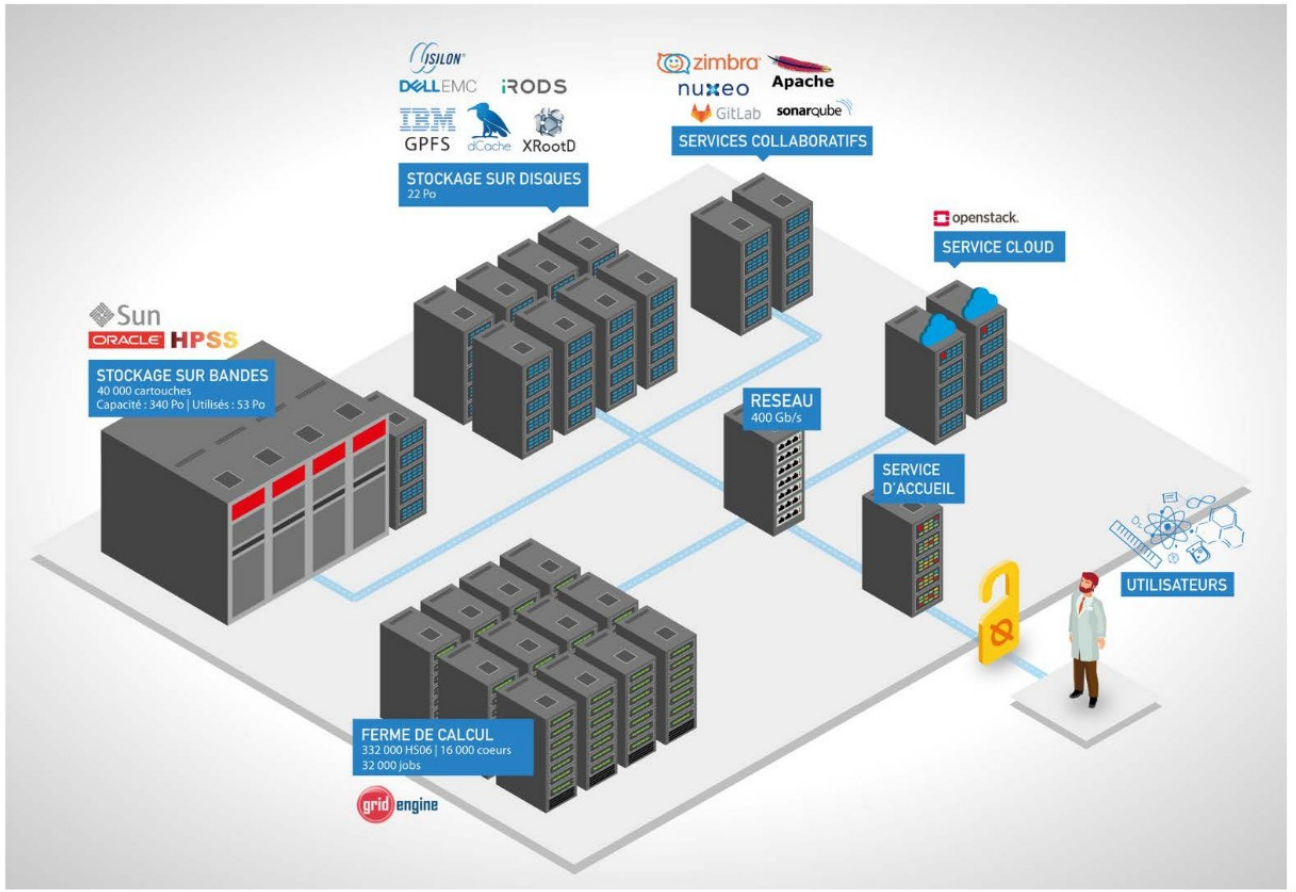
Ressources humaines: 80 (65 ingénieurs)
1700 m2 en 2 salles informatiques
Calcul: 850 serveurs de calcul
Consommation d'électricité: **1,4 MW**
Stockage: **276 Po** **(60 % bande magnétique)**

Infrastructure: sockage sur bande

Bandothèques automatisées

Espace utilisé: **171 Po**
Capacité disponible: **360 Po**
Pas de préservation/archivage

Infrastructure pour construire services



Infrastructure

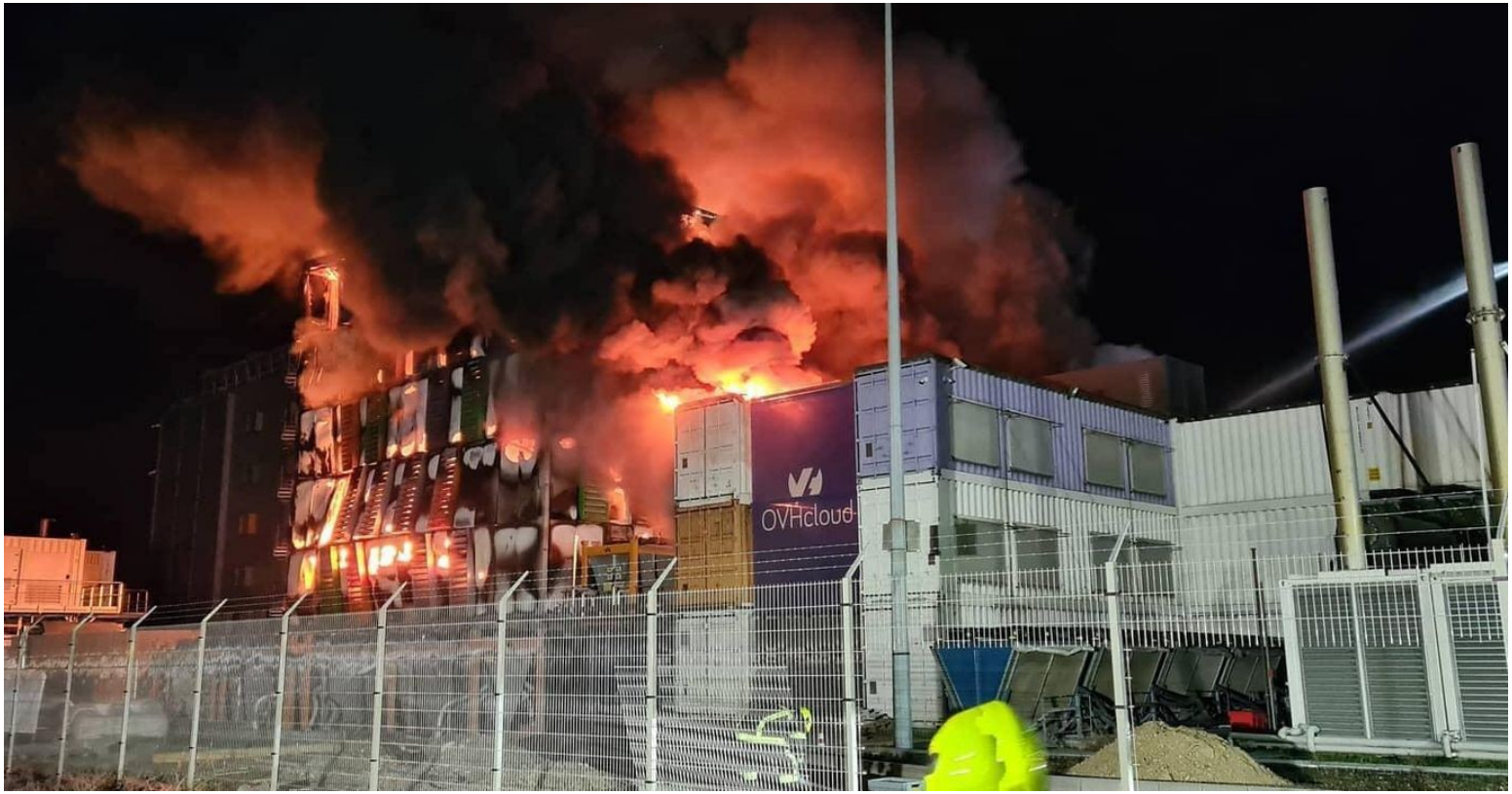
- Organisation
  - Équipe des personnes/compétences
  - Procédures d'opération
  - Gestion des risques
- Ressources informatiques
  - Entretien et rénovation parc
  - Évolutions technologiques et besoins
  - Gestion des incidents et dépannages
  - Optimisation des coûts
- Sécurisation et protection locaux
- Niveau de disponibilité: 24h/24h – 365 jours
  - redondance électrique, réseau, …
- Cybersécurité
- Migration de données
- Tous ces aspects sur le long terme

Clé: mutualisation

Infrastructure: le « dual site »

L'incendie du centre de données d'OVHcloud à Strasbourg le 10 mars 2021

- Les données contenues dans les serveurs détruits sont perdues
- « Le 26 janvier 2023, le tribunal de commerce de Lille rend un jugement qui établit qu'OVHcloud a commis un manquement contractuel en stockant les sauvegardes au même endroit que les données principales et qui condamne la société OVHcloud à payer dommages et intérêts »
- Ce site de Strasbourg ne comportait pas de dual site
- Les sinistrés, affectés à différents degrés incluent notamment le site de data.gouv.fr



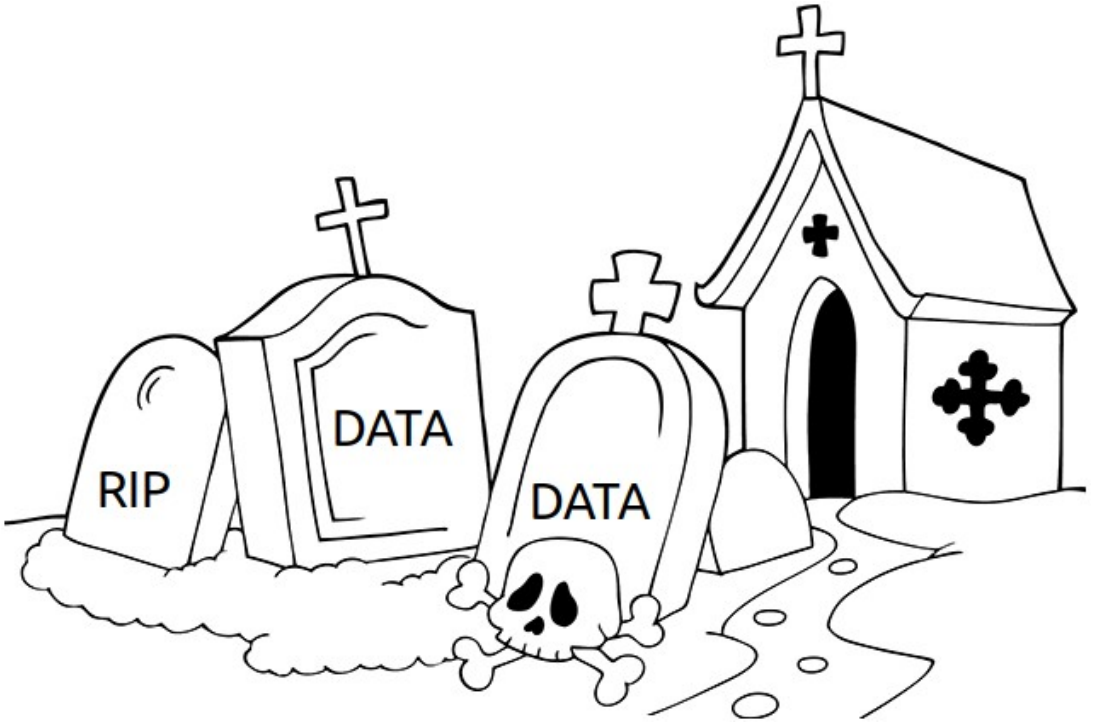
Motivation

Les données scientifiques au CC-IN2P3

- Constat en 2013
- Accumulation de données historiques (plus de 30 ans)
  - Données orphelines
    - (absence de propriétaire identifié joignable)
  - Impossible d'établir l'intérêt
  - Impossible de prendre des décisions (retenir, détruire, …)
- Absence de catalogue de l'ensemble des données stockées
- Absence d'une politique claire et prédéfinie concernant le cycle de vie des données hébergées

Motivation

Les données scientifiques au CC-IN2P3



- Proportion des données mortes difficile à établir
- Interdiction de fouiller ou détruire les données !
- On doit continuer à les stocker et les gérer comme des données actives

Motivation

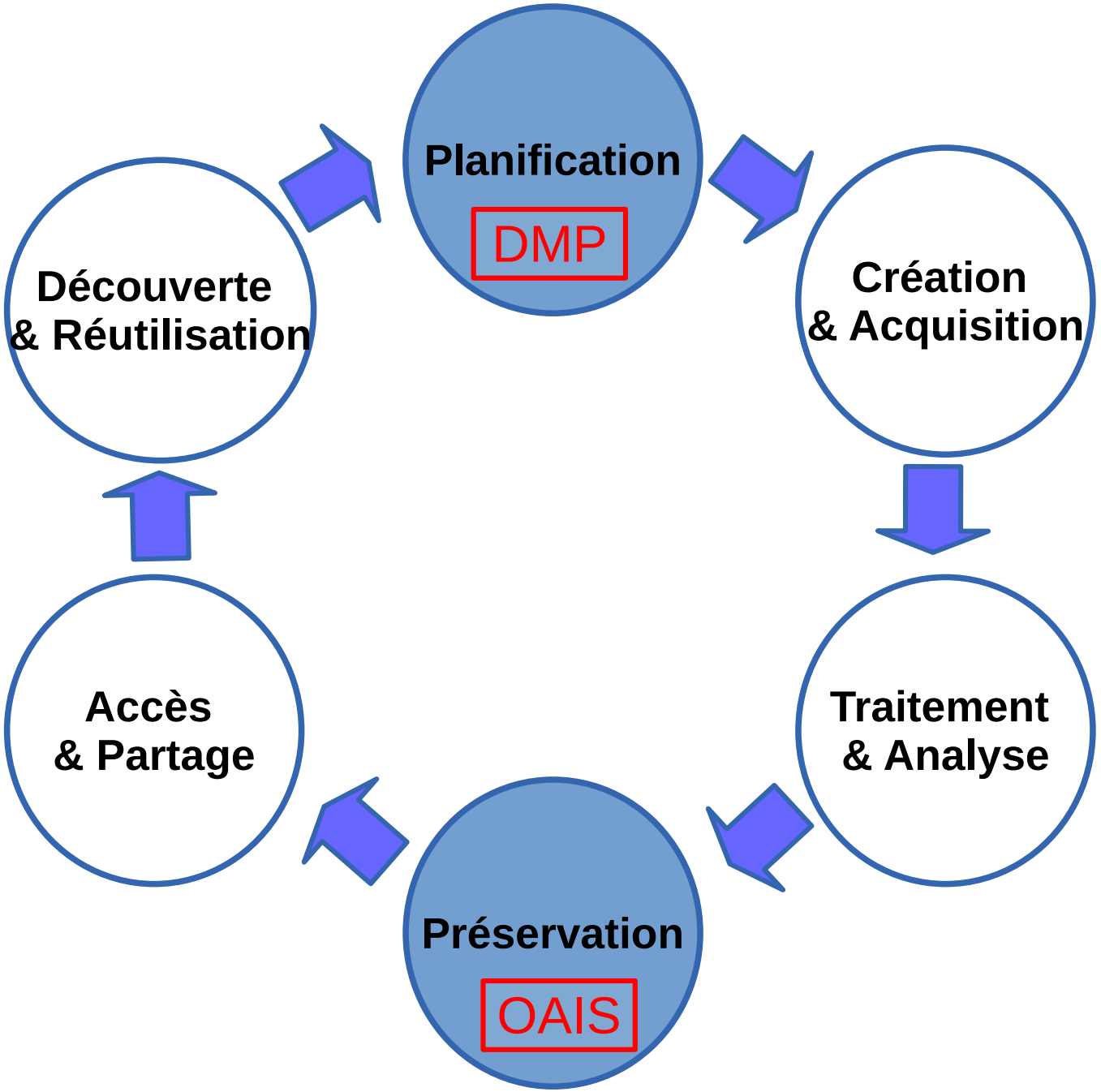
Les données scientifiques au CC-IN2P3

- Quelle contribution depuis l'infrastructure ?
- Groupe de travail interne créé en 2014
  - Devenir des Données (3D)
- Données orphelines: rien ne peut être fait – **R.I.P. DATA** (archéologie?)
- Contenir le problème dans le futur

Propositions du groupe de travail 3D:

- Création d'un catalogue du contenu de l'ensemble des systèmes
- Développement du Plan de Gestion de Données
- Étude de faisabilité pour l'implémentation du service de préservation de données à long terme en suivant le modèle OAIS

Gestion cycle de vie des données



Étude de faisabilité

- Réaliser l'état de l'art :
  - Littérature: modèle OAIS,CSIP/DILCIS Board, …
  - Bonnes pratiques de préservation: iPRES , …
  - Organisations de référence: BnF, CINES, …

- Structuration de la notion de «service de préservation» :
  - Adaptation au contexte de recherche IN2P3/CNRS
  - Cahier des charges, 41 pages (rendu en juillet 2020)
  - Mise en place du service pilote (Data Future)
    - indices ou notions permettant d'estimer complexité, coûts,…

Principes de base de Data Future

- Projet pilote d'un service de preservation
- Dissociation en deux activités principales:
  - Infrastructure (préservation de l'octet)
  - Curation et valorisation (préservation de l'information)

- Action volontaire de la part du producteur de la donnée (chercheurs, projet ou expérience scientifique)
- Aller à l'essentiel, le plus simple et flexible possible
- Principal objectif la réutilisation des données: s'ancrer dans le contexte de science ouverte
- Utiliser l'infrastructure existent
- Destiné aux nouveaux projets de recherche
  - «préserver les données avant de les produire»

Your present data for the future

https://irods.in2p3.fr/datafuture

Data Future - caractéristiques principales

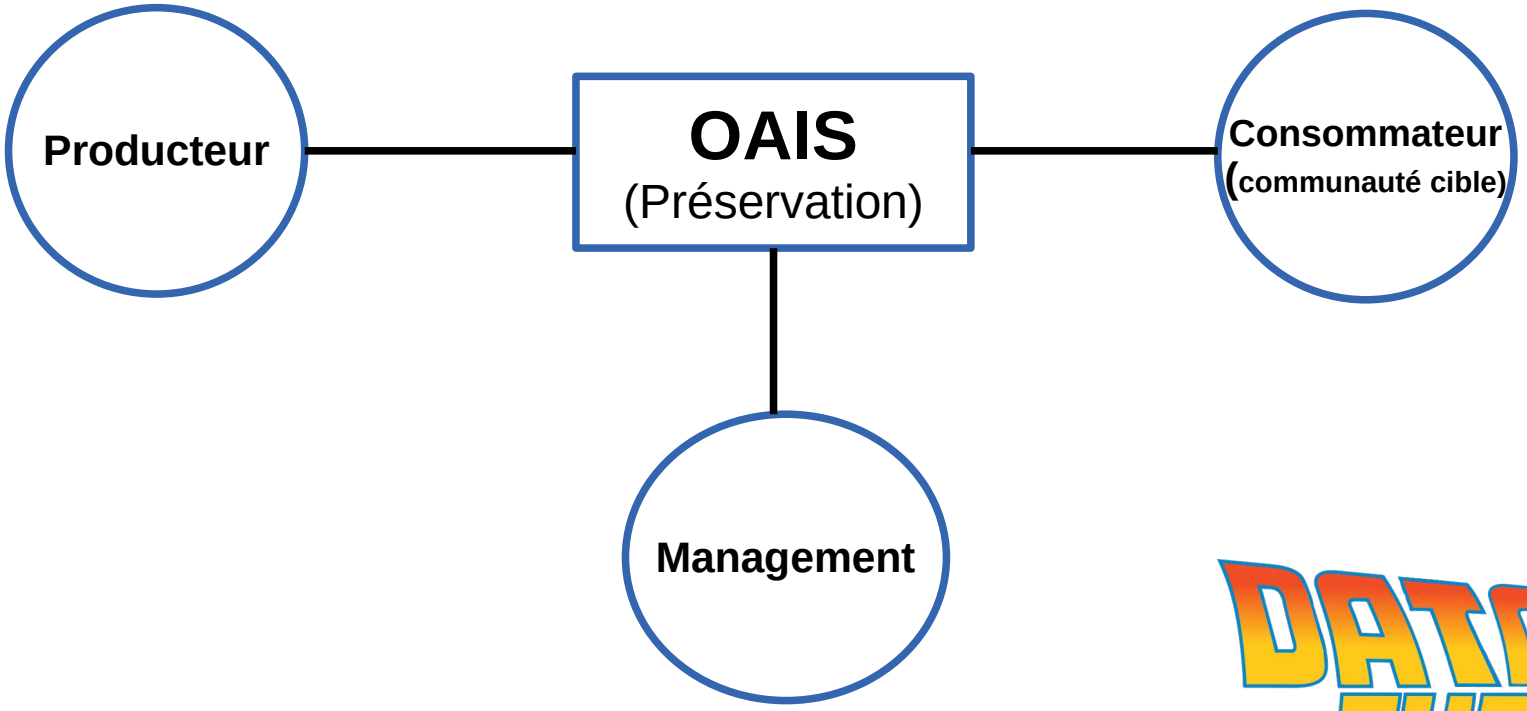
- Simple et flexible pour les producteurs de données
- Conçu pour prendre en charge de grandes quantités de données scientifiques
- Responsable du contenu des données déposées
- Périodes de temps courtes ou longues (années - décennies)
- Complément au processus F.A.I.R et Science Ouverte
- Applique les normes OAIS et les meilleures pratiques de préservation numérique
- Permet la diffusion personnalisée de données

Ce que Data Future ne fait pas:

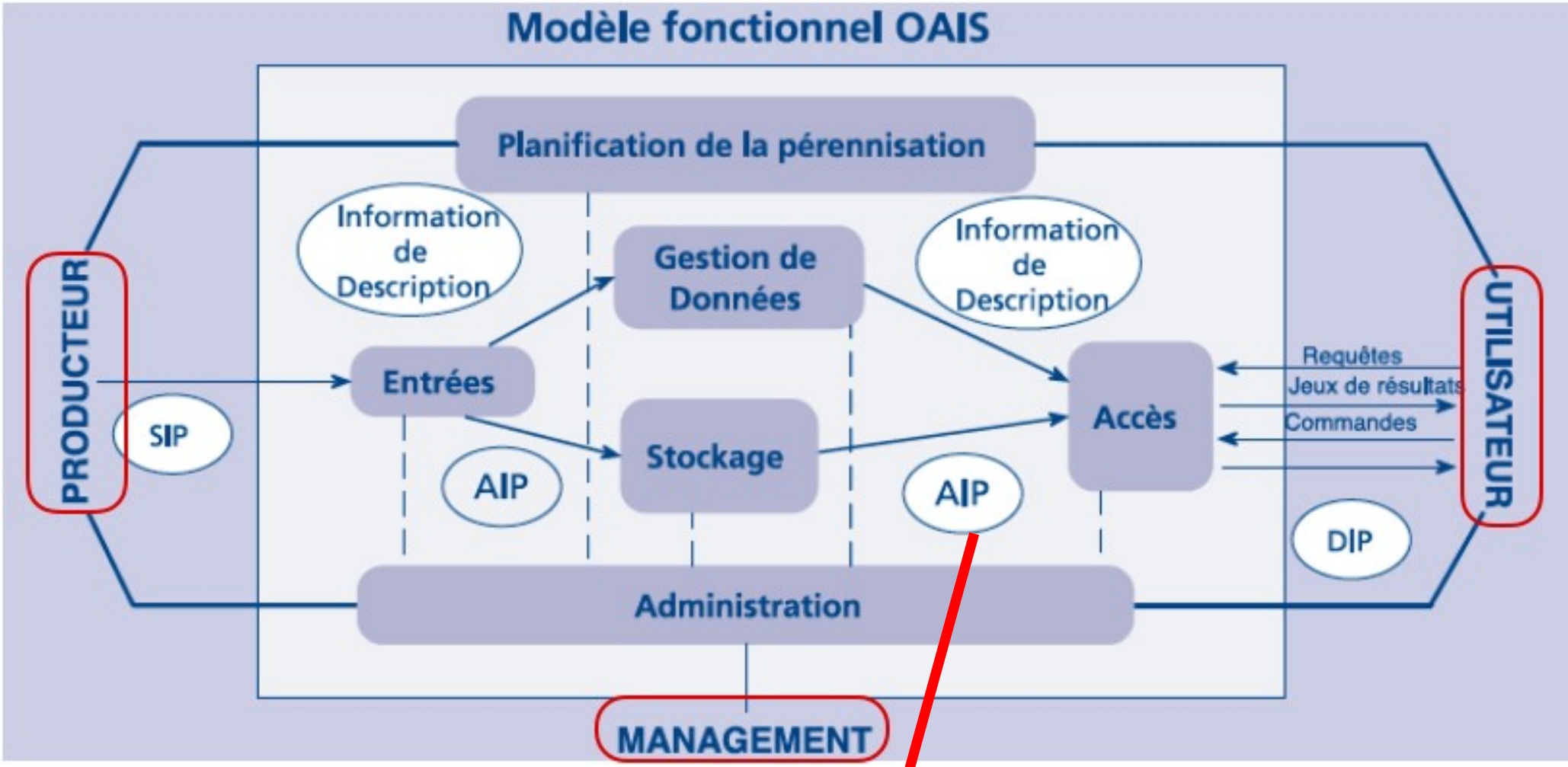
- Curation et valorisation des données
  - application des principes F.A.I.R aux données
  - fait par le producteur de données
    - Chercheur
    - Projet ou expérience scientifique
    - Intermédiaire (e.g. dépôt de données, …)
- Dépôt légal de documents numériques administratifs
- Valeur probante
- Conservation ad vitam aeternam

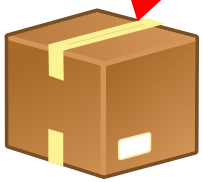
OAIS: Système Ouvert d'Archivage d'Information

L'OAIS spécifie de manière très générale l'architecture logique et les fonctionnalités d'un système de préservation numérique (archivage). **C'est un modèle abstrait qui définit une terminologie et des concepts**. Il identifie les acteurs, décrit les fonctions et les flux d'information et propose un modèle d'information particulièrement adapté à la problématique de la préservation numérique, tout cela indépendamment de la nature des objets à conserver.



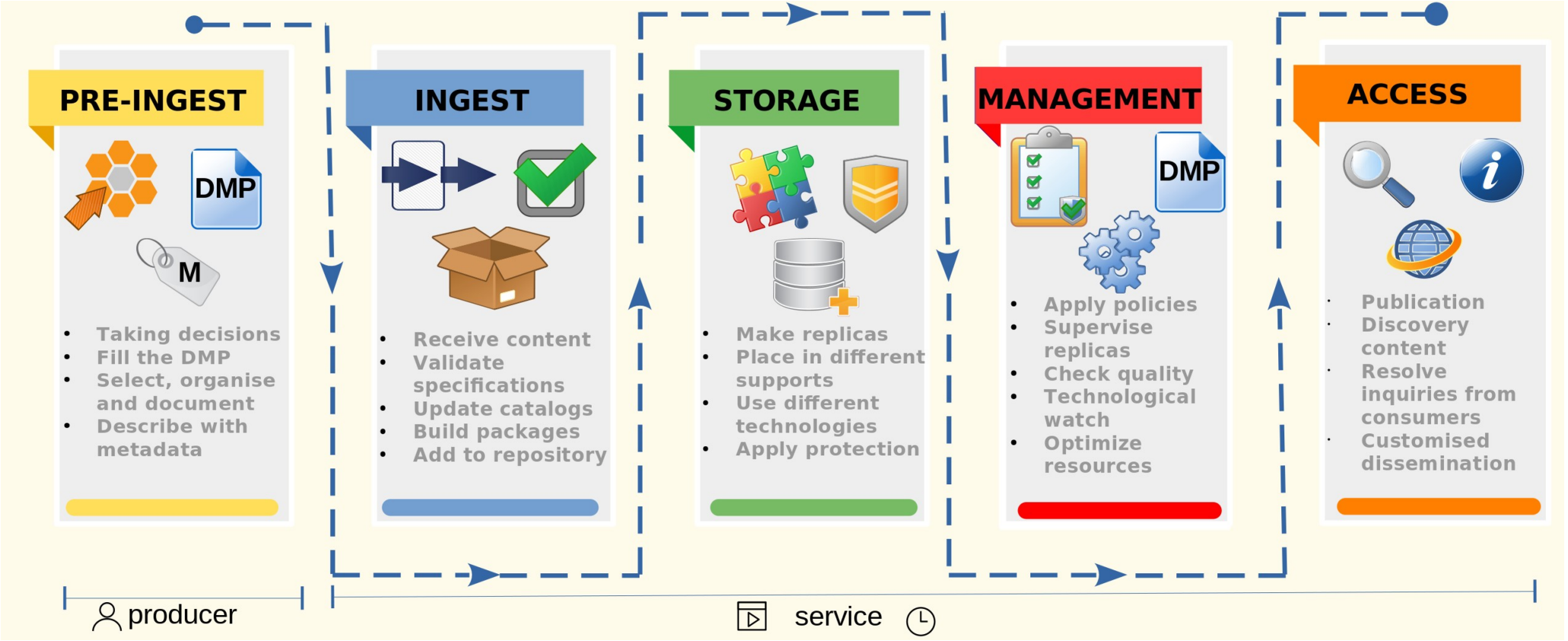
OAIS: modèle fonctionnel





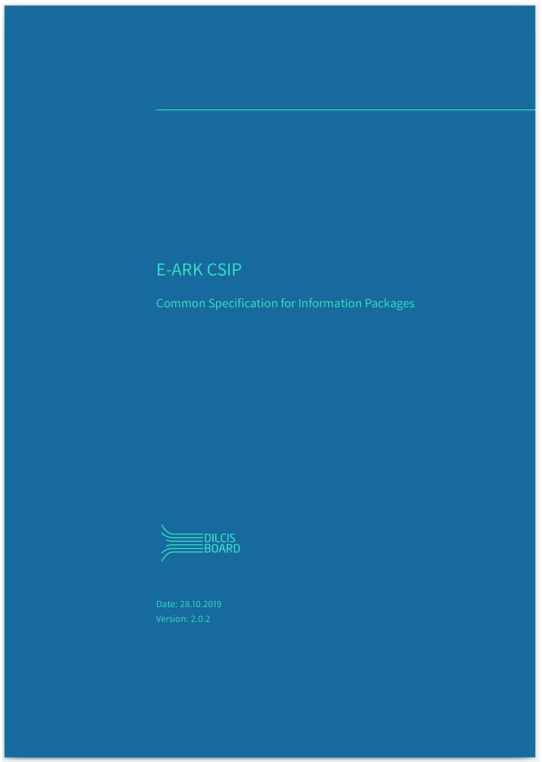
Paquet d'Informations Archivé (AIP)

Processus de préservation



Common Specification for Information Packages

CSIP ou E-ARK Specifications



Objectifs

- Automatisation de l'identification et de la validation de paquets d'information (OAIS).
  ➔ intégrité, validité technique,…

- Séparer les composants d'un paquet d'information
  ➔ données, métadonnées, documentation

- IP doit être à la fois lisible par l'homme et exploitable par la machine

- Implémentation basée sur le standard de métadonnées **METS**

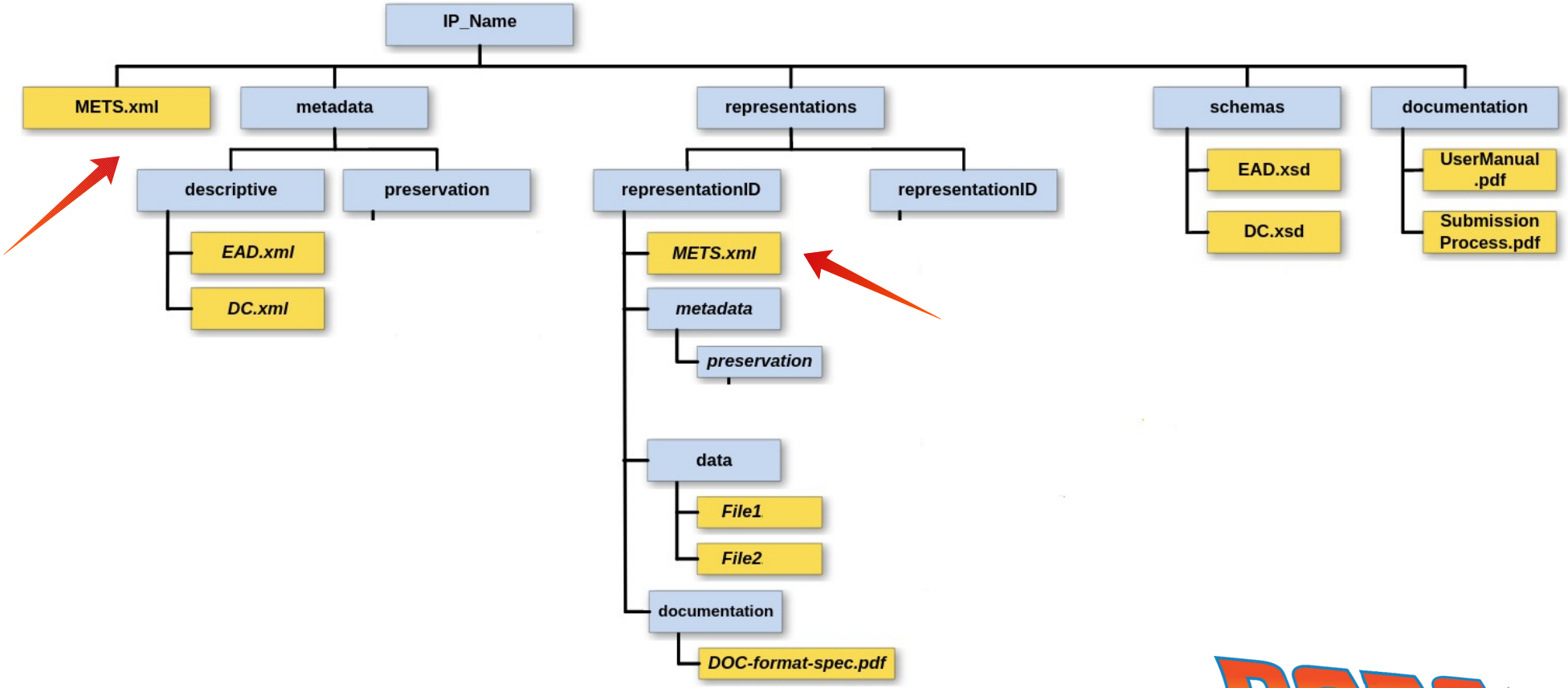
DILCIS BOARD

DLM FORUM

DATA FUTURE IN2P3

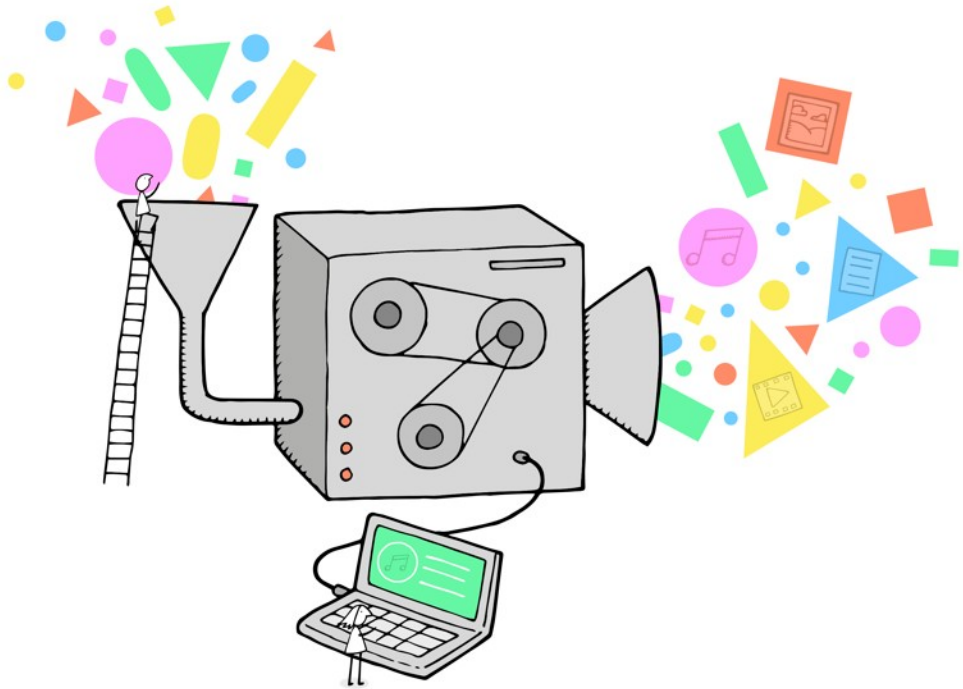
Common Specification for Information Packages

CISP information package structure



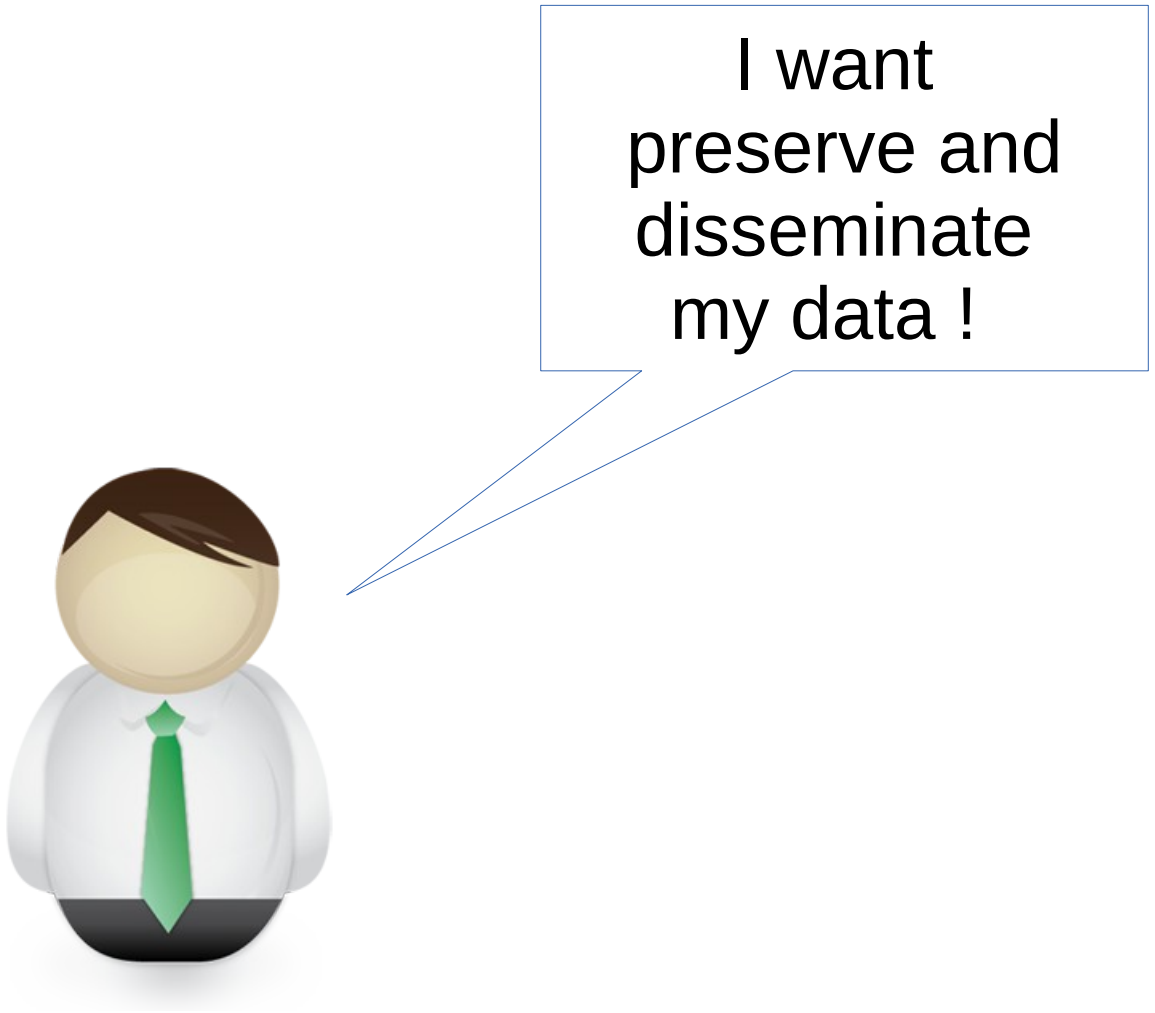
C’est quoi la Préservation Numérique ?

Série d'activités gérées (effort formel) pour garantir que les informations numériques restent **accessibles** et **réutilisables** sur le long terme

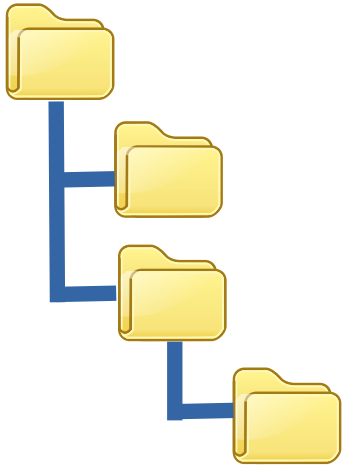


Digital Preservation Business Case Toolkit http://wiki.dpconline.org/, CC-BY-NC 3.0

Producteur de données: chercheur



Producteur de données: metadata et DMP

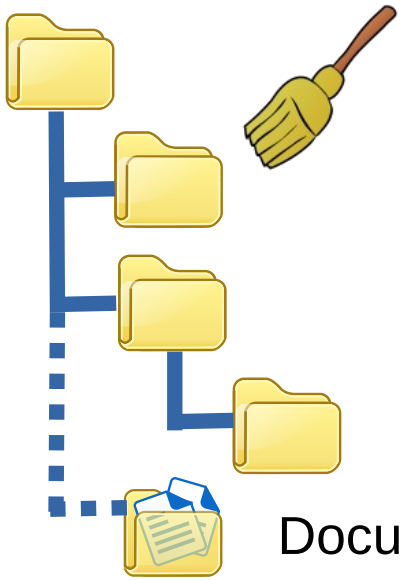


DATA

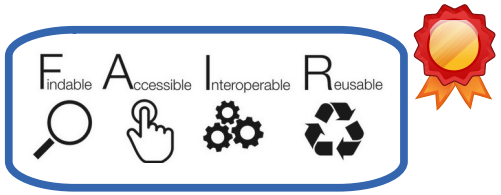
Metadata

DMP

Gestion des données de la recherche



DATA

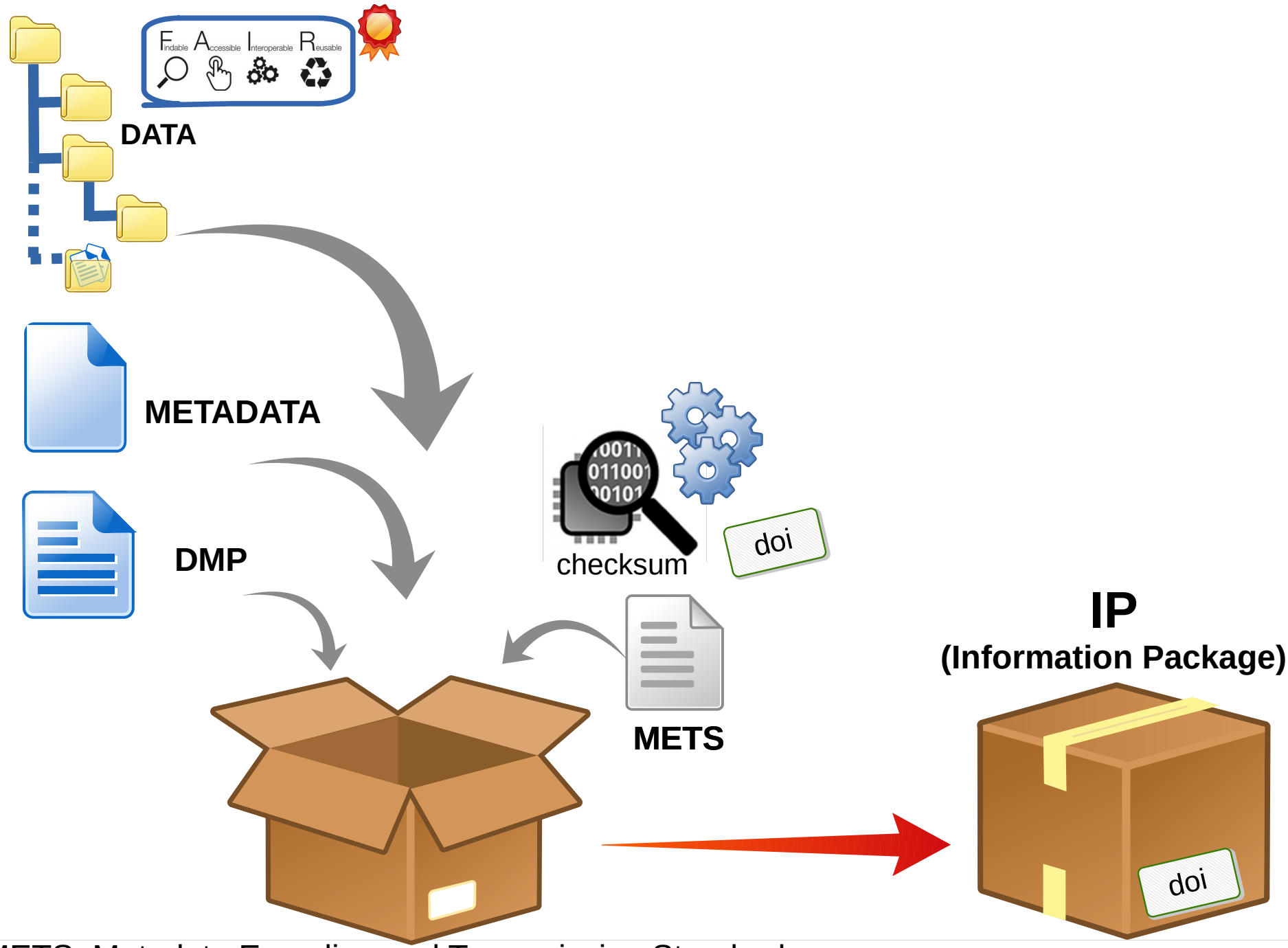


Documentation

Metadata

DMP

Paquet d'Informations



METS: Metadata Encoding and Transmission Standard

Data Future: Réalisations (1/2)

- Étude de faisabilité
  - Service de préservation avec infrastructure existante
- Cahier des charges en 2020 pour la description du service de préservation
  - "Validé" par entités externes (ENSIB et AAF)
  - Plan d'action -- estimation des coûts
- Implémentation du Common Specification for Information Packages
  - Jeux de données == Information Package
  - Grande volumétrie: centaines de To et milliers de fichiers
- Développement du logiciel createAIP basé sur la bibliothèque logiciel RODA
  - Création automatique des paquets (SIP et AIP) depuis les serveurs de données
  - Segmentation des AIPs

Data Future: Réalisations (2/2)

- Intégration avec iRODS
  - iRODS conception fortement influencé par la communauté de préservation
  - Validation automatique du processus de versement (SIPs)
  - Gestion de AIPs en bande magnétique (double copie)
  - Gestion de métadonnées pour le service
  - Accès et diffusion via web

- Interopérabilité avec DMP machine actionnable
  - Questionnaire "universel" de 123 questions organisées en 7 sections et 42 sous-sections
  - Plateforme logiciel de gestion et collaboration basée sur RDMO https://dmp.in2p3.fr
  - DMP est transformé et normalisé (maDMP RDA) en métadonnées administratives

- Sensibilisation
  - Conseil et accompagnement auprès des expériences et projets
  - Formations écoles doctorales - Université de Lyon

Data Future: Utilisation (1/4)

- Expérience: **EROS - Expérience de Recherche d'Objets Sombres**
- Domaine: Astrophysique et Astronomie
- Partenaires: CEA-IRFU, IN2P3/CNRS, CDS/CNRS
- Sujet: Recherche et étude des corps stellaires sombres, appelés «naines brunes» dans les nuages de Magellan (deux galaxies naines en orbite de la Voie Lactée) sur plus de 10 ans, du début des années 1990 jusqu'en 2003
- Jeux de données : **46 To - 106 millions fichiers**
- Préservation: **7301 segments AIPs (pre-production)**
- État actuel: Finalisation du processus de curation
- Observations:
  - projet d'archéologie - Eros Anastasis https://groups.ijclab.in2p3.fr/erosanastasis
- Site web: http://eros.in2p3.fr

Data Future: Utilisation (2/4)

- Expérience:**Vigie-Chiro - Programme national de suivi des chauves-souris**
- Domaine: Sciences de la vie, Zoologie, Écologie
- Partenaires: MNHN (Museum National d'Histoire Naturelle), INEE/CNRS
- Sujet: Programme de suivi des populations de chauves-souris basé sur les sciences participatives et les enregistrements acoustiques permettant l'identification et l'étude des chiroptères.
- Jeux de données: **456 To - 1.4 millions fichiers**
- Préservation: **151786 segments AIPs (production) - 284To (62%)**
- État actuel:
  - en production, données historiques (2014-2022),
  - flux de données actuelles (de 2022) à faire
- Observations:
  - l'expérience manipule directement les AIPs
  - volumétrie actuelle 1 Po (double copie en bande magnétique)
- Site web: https://www.vigienature.fr/fr/chauves-souris

Data Future: Utilisation (3/4)

- Expérience: **IPM-DIAM, Dispositif d'Irradiation d'Agrégats de Molécules à l'IP2I**
- Domaine: Physique, Chimie, Physique Chimie Atmosphérique, Astrochimie, Spectrométrie de masse, Chimie froide, Science des rayonnements
- Partenaires: IN2P3/CNRS, Université Claude Bernard - Lyon, INSU/CNRS
- Sujet: Étude de l'irradiation dans des nanosystèmes moléculaires avec deux axes principaux:
  - les enjeux dans l'atmosphère terrestre dans le contexte des études sur les changements climatiques
  - les enjeux sur l'origine des molécules du vivant dans le contexte astrophysique
- Jeux de données: **12 To - 8 millions fichiers**
- État actuel: en attente de la résolution des questions juridiques
- Observations:
  - historique (1990-2020) prêt
  - flux de données actuelles (de 2024) déjà prêt à la conservation et à la curation en temps réel
  - conservation de toutes les données (acquisition, cahier de laboratoire, échanges internes,…)
- Site web: https://www.ip2i.in2p3.fr/equipes/ipm

Data Future: Utilisation (4/4)

- Expérience: **Expérience: GRAND (Giant Radio Array for Neutrino Detection)**
- Domaine: Sciences Naturelles / Astrophysique et Astronomie
- Partenaires: (collaboration internationale) ANR, CNRS, Radbout University (Nederlands), Karlrushe Institut of Technology (Germany), Penn State University (USA), IIHE, Bruxelles (Belgium), National Astronomical Observatories of China (China), …
- Sujet: Réseaux d'antennes radio sur des centaines de kilomètres carré dans la pampa argentine pour la détection de neutrinos de ultra-haute énergie
- Jeux de données: **2.6 Po (attendue avant 2030)**
- Préservation: En planification
- Observations:
  - Pas de données produites, préservation et diffusion en conception
  - Préservation à source de la production de données
  - Diffusion (Open Data) jeux de données
- Site web: https://grand.cnrs.fr

Au-delà de la préservation de données

Comment réutiliser (exploiter) les données dans le futur?

Environnement système + données + logiciels

- Émulation
  - matériel et logiciels qui utilisent les données
  - stratégie de préservation
- Objectifs
  - conserver et reproduire l'environnement fonctionnel d'origine
  - pouvoir continuer à l'exécuter à long terme
- Encapsulation
  - données + logiciels (métier)
- Virtualisation
  - matériel + logiciels (système)
- Challenge
  - suivi et validation des émulateurs dans le temps

Quick EMUlator

Résumé

- L’infrastructure est indispensable pour la préservation numérique mais elle n’est pas suffisante
- RIP DATA a un très fort impact mais difficile à quantifier
- La mutualisation de l’infrastructure est un aspect clé pour le financement
- La préservation de données scientifiques demande de la rigueur avec l’application de standards et normes
- Les données scientifiques doivent être préservées avant d'être produites
- Il faut développer la culture de la gestion du cycle de vie de données à tous les niveaux
- Il faut créer un écosystème numérique pour la préservation (centres de données, entrepôts, …)
- La préservation numérique doit faire partie intégrante de la Science Ouverte
- Manque de volonté politique. Le CNRS et ses instituts doivent définir leurs ambitions: le caractère obligatoire, leurs missions, leurs recommandations,… ne sont pas effectifs

Remerciements

- **Céline Guyon**
  Ancienne présidente de l'Association des Archivistes Français (AAF)
- **Thomas Ledoux**
  Ancien chef de projet SPAR (Système de Préservation et d'Archivage Réparti)
  Bibliothèque Nationale de France (BNF)
- **Laurent Duplouy**
  Chef du service multimédias du département de l’audiovisuel
  Bibliothèque Nationale de France (BNF)
- **Lorène Béchard**
  Ancienne responsable fonctionnelle du système d'archivage électronique
  CINES
- Équipe stockage du CC-IN2P3

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

Data Future – Les données du présent pour le futur

Yonny CARDENAS

cardenas at cc.in2p3.fr

RIP Data: quelle sélection, conservation et suppression des données de recherche?

1-2 octobre 2024, Aix-en-Provence

Archivage des données de la recherche

Décrets, arrêtés, circulaires

TEXTES GÉNÉRAUX

MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Arrêté du 29 avril 2016 relatif à l'Institut national de physique nucléaire et de physique des particules du Centre national de la recherche scientifique

NOR : *MENR1611333A*

La ministre de l'éducation nationale, de l'enseignement supérieur et de la recherche,

Vu le décret n° 82-993 du 24 novembre 1982 modifié portant organisation et fonctionnement du Centre national de la recherche scientifique, notamment ses articles 12, 14 et 16 ;

Vu la décision du président du Centre national de la recherche scientifique du 21 janvier 2010 modifiée portant création et organisation des instituts et fixant la liste des sections et des commissions interdisciplinaires concernées par leur activité,

Arrête :

Art. 1er. – L'Institut national de physique nucléaire et de physique des particules du Centre national de la recherche scientifique exerce les missions nationales d'animation et de coordination dans les domaines de la physique nucléaire, de la physique des particules et des astroparticules, des développements technologiques et des applications associées, notamment dans le champ de la santé et de l'énergie, en ce compris la radiochimie.

Pour la réalisation de ces missions, l'Institut national de physique nucléaire et de physique des particules :

- conçoit, coordonne et anime des programmes de recherche nationaux et internationaux dans ses domaines de compétence ;
- organise et conduit, en y associant les organismes et acteurs concernés, des exercices de prospective nationale permettant de définir la stratégie scientifique de long terme et d'identifier les équipements nationaux et internationaux nécessaires à sa mise en œuvre. Il veille à la plus large diffusion des résultats de ces travaux et favorise leur prise en compte dans l'élaboration des programmes de recherche et d'équipement à l'échelle nationale et internationale ;
- favorise et coordonne la participation des opérateurs de recherche aux structures d'intérêt national ainsi qu'aux très grandes infrastructures de recherche et aux programmes scientifiques qu'elles permettent de réaliser ;
- coordonne la mise en place de systèmes d'information permettant le stockage, la mise à disposition auprès de la communauté scientifique, le traitement et la valorisation de l'ensemble des données scientifiques concernées, ainsi que leur archivage.

Art. 2. – Un conseil d'orientation est créé au sein de l'Institut national de physique nucléaire et de physique des particules du Centre national de la recherche scientifique.

Archivage des données de la recherche

Arrêté du 29 avril 2016 relatif à l'**IN2P3**

Missions

*"- Coordonne la mise en place de systèmes d'information permettant le stockage, la mise à disposition auprès de la communauté scientifique,le traitement et la **valorisation** de l'ensemble des données scientifiques concernées, ainsi que leur **archivage**."*

Journal officiel de la République française - N° 122 du 27 mai 2016

AIP Segmentation avec CSIP

- Manipulation des AIP de grand volume

- Par défaut, un AIP est censée résider dans un seul dossier ou fichier
- Un seul AIP peut facilement atteindre plusieurs To et devenir difficile à gérer
- Les gros AIPs peuvent être divisés (segmentés)
- Le CSIP peut être étendue pour prendre en charge la segmentation
- Un segment parent et plusieurs segments fils
  - Description en METS