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

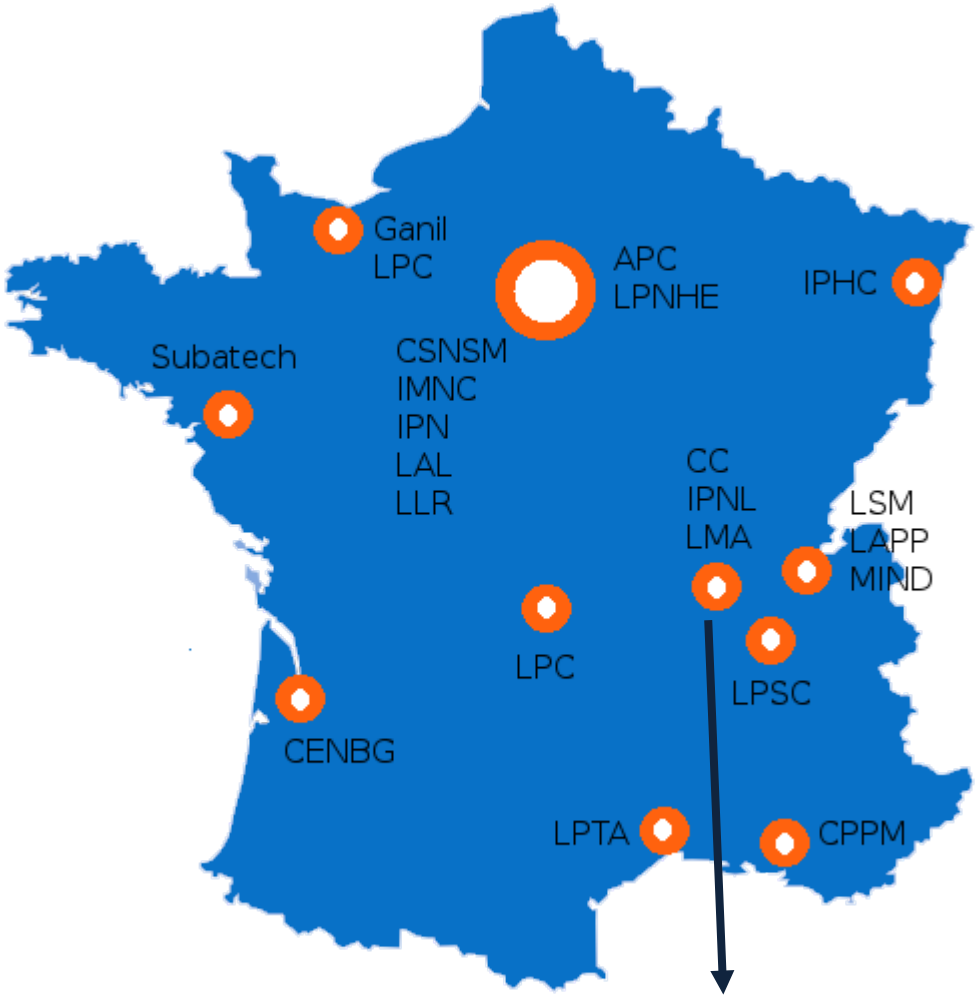
# Archivage des Données à l'IN2P3

Yonny CARDENAS

Journée archivage numérique des données de recherche

Grenoble, 20 Novembre 2019

# Le CC-IN2P3





IN2P3:
Institut National de Physique Nucléaire et Physique des Particules
L'un des 10 instituts du CNRS
Composé de 20 laboratoires
Domaines scientifiques
– Physique des particules
– Physique nucléaire
– Physique des astroparticules

CC-IN2P3:
Centre de Calcul de l'IN2P3
– installé à Lyon depuis 1986

La mission du CCIN2P3 est de fournir des services informatiques aux laboratoires de l'IN2P3
– Calcul
– Stockage de masse
– Service web, Bases de données
– Outils collaboratifs

# Le CC-IN2P3

- CC-IN2P3 propose:
  - Stockage et  ressources de calcul
    - Local, grid et cloud pour l'accès
  - Service Base de données
  - Hébergement de sites web

- 2100 utilisateurs actifs (plus avec grid):
  - 600 utilisateurs à l'étranger .

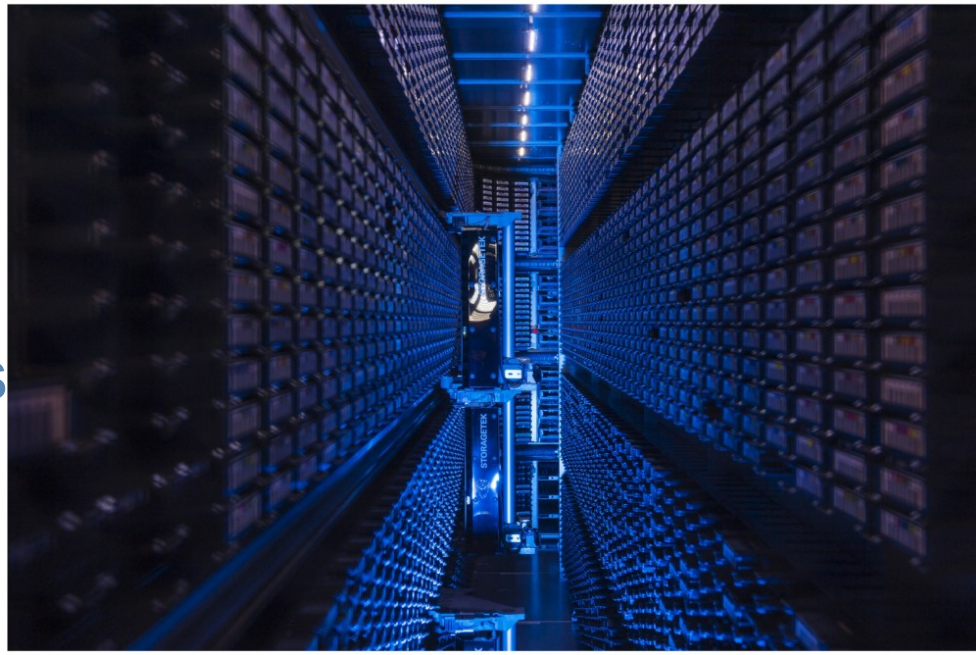
- ~ 140  groupes actifs (laboratoires, expérience, projet).

- ~ 40000 cores système batch

- ~ 100 PBs des données stockées en disque bande magnétique.

# Les Systèmes de Stockage

- Stockage de masse
  - Bandes magnétiques et disques pour le cache
    - utilisées par **HPSS** dans des bibliothèques
- Stockage disque
  - DAS (Direct Attached Storage)
    - utilisé par **dCache, XRootD, iRODS**
    - frontal pour l'accès au stockage de masse
  - NAS (Network Attached Storage)
    - utilisé par Isilon
    - accès POSIX via montage **NFS**
  - SDS (Shared Disk Storage)
    - utilisé par **GPFS** accès POSIX



# Les Données au CC-IN2P3

Diversité des domaines scientifiques

Deux grandes catégories :

- Physique nucléaire et des particules et  Astrophysique (~ 90 %)

- Ouverture interdisciplinaire :
  - Neurosciences
  - Biologie
  - Médecine
  - Informatique
  - Chimie
  - Écologie
  - Sciences humaines et sociales (Huma-Num)

# Les Données au CC-IN2P3

Diversité des types de fichiers

- Très grande variété de format
- Raw data (appareil de mesure, relevé de terrain)
- Banques de données communautaires
- Simulation
- Analyse

Diversité des technologies de stockage:

| | |
|---|---|
| GPFS | HPSS |
| IRODS | XROOTD |
| DCACHE | ORACLE |
| TSM | MYSQL |
| NFS | POSTGRES |

# Politique Gestion des Données en vigueur

- Migration des données sur des supports récents
  - Préservation des octets
- Les données sont accessibles pendant la durée du projet
- Pas d'effacement systématique des données y compris en fin de projet
- Certaines zones de stockage peuvent être sauvegardées à la demande
- Désignation d’un responsable des données par projet
- Respect des engagements du MoU (Memorandum of understanding)

# Politique Gestion des Données en vigueur

## Qualités

- Données accessibles en temps réel
- Possibilité de relire les données anciennes sur des médias récents
- Investissement de temps minimal pour l'utilisateur concernant la gestion des données

## Faiblesses

- Difficulté d'identifier et de valoriser les données précieuses
- Données temporaires ou orphelines pas supprimées entièrement
- Données vieillissantes pouvant devenir inexploitables
- Détournement des services pour pérennisation de données
  - e.g. service de sauvegarde
- Pas de garantie de réutilisation des données

*Nécessité de la mise en œuvre d'une plateforme d'archivage*

# Service d’Archivage CC-IN2P3

Projet : Mise en place d’un service d'archivage intermédiaire

S’adresse aux jeux de données qui ont cessé d’être considérés actifs et peuvent faire l’objet de sélection pour préservation à long terme.

Un service se plaçant entre le stockage temporaire et la conservation définitive et en appliquant une partie des méthodes de préservation numérique à long terme.

Stratégie de migration de format n'est pas viable

- Diversité de formats et d’utilisations

# Service d'Archivage CC-IN2P3

Projet : en phase de prototypage

Creation de Paquet d'Information SIP AIP DIP

- OAIS
- Common Specification for Information Packages - E-ARK CSIP
- European Archival Records and Knowledge Preservation Project (E-ARK project)
- IP manipulation java library project RODA
- Test stockage utilisant Linear Tape File System (LTFS)
- Assignation DOI
- Construction  de catalogue basé sur iRODS
- Diffusion avec fonctionnalités  iRODS
  - Authentification
  - Métadonnées
  - API (REST, java, C++, …)

# Open Data & Archivage

CHERCHEUR

DOCUMENTALISTE

INFORMATICIEN

ORGANISER

DOCUMENTER

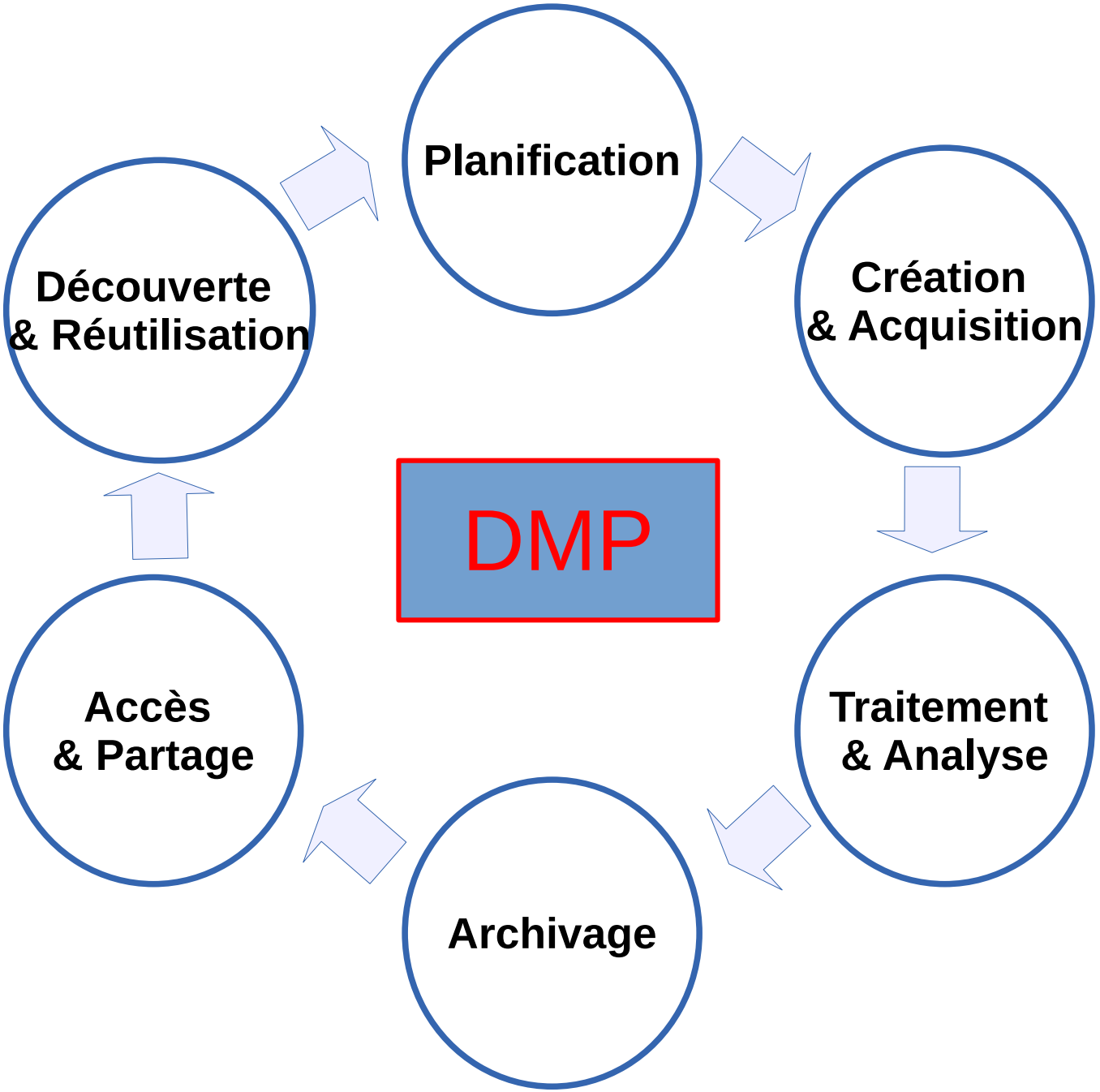
**DATASET**
**(JEUX DE DONNÉES)**

MÉTADONNÉES

# Paquet d'Information Archivé (AIP)

# Préservation & Découverte

# Cycle de vie des données de recherche

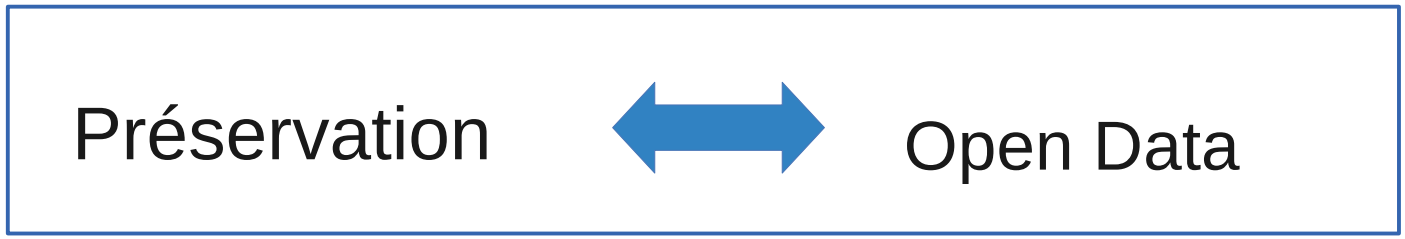


# Archivage & Science Ouverte

Objectif : Réutilisation de l'information

Quelques processus en commun à mutualiser :

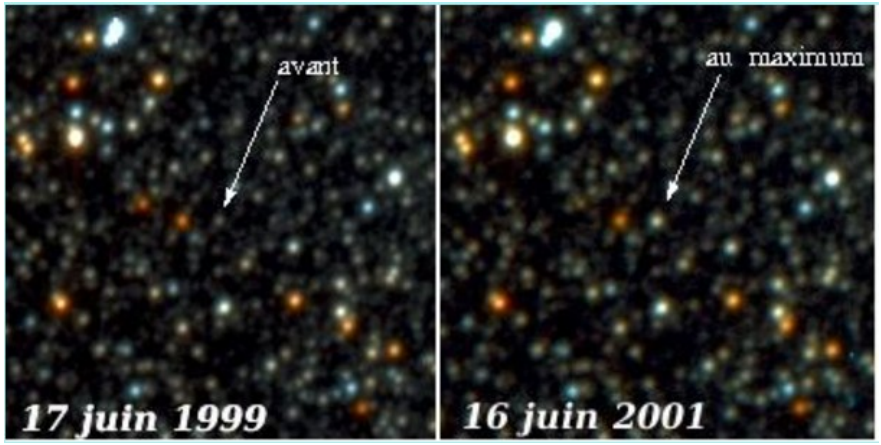
- Organisation et structuration de données
- Documentation
- Description (Métadonnées)
- Planification (Plan de Gestion des Données)



# Cas d'utilisation

## EROS Experiment

- « EROS: Expérience pour la Recherche d'Objets Sombres »
  - Phases observationnelles:
    - EROS-1 (1990-1995)
    - EROS-2 (1996-2003)
  - Astroparticules
    - Matière noire et énergie sombre
    - Mettre en évidence l'existence d'étoiles naines, trop petites pour être lumineuses, mais qui participeraient à cette masse de l'Univers.
    - Observatoire européen austral à La Silla au Chili
    - http://eros.in2p3.fr




# Cas d’utilisation

EROS Experiment

- Données:
  - Type de fichiers:  images en format FITS
  - Stockées sur bande magnétique : HPSS
  - Volumétrie: 27 TiB
  - Nombre de fichiers:  412 mil
  - Plusieurs années sans activité (accès)
  - Plusieurs migrations (support stockage)

- Fort intérêt de la communauté internationale en Astroparticules
  - Valeur scientifique (date de prise)
  - Constat: disparition de compétences (individus)
  - La crainte de perdre l’information

# Cas d’utilisation

EROS Experiment

- Projet de « résurrection des données Eros »
  - Valorisation des données de l’expérience  stockées au CC-IN2P3
  - équipe d’ingénieurs et de physiciens
  - Laboratoire de l'Accélérateur Linéaire - LAL-IN2P3 Orsay
  - Démarré depuis Novembre 2017

- Objectifs initiaux (9 Novembre 2017):
  - Créer un catalogue de fichiers
  - Croiser ce catalogue avec le catalogue des images et des courbes de lumière de la base de données Oracle.
  - Mettre en place des outils, éventuellement simples, pour accéder aux fichiers.

# Cas d'utilisation

EROS Experiment

- Pourquoi il est intéressant pour l'archivage
  - Intérêt explicite d'une communauté cible
  - Objectif: le rendre réutilisable à long terme
  - Objectif de diffusion large (e.g. observatoire virtuel) - Science Ouverte
  - Volumétrie relativement importante (pas d'offre d'archivage)
  - Format de données spécifique
  - Travail interdisciplinaire
  - Exemple de mutualisation: démarches science ouverte et archivage
    - Curation, Organisation,
    - Documentation, Description
    - Génération de Métadonnées
    - Diffusion

# Conclusion

- Besoin réel d'archivage à long terme de jeux de données scientifiques
- Nouvelle activité compétence multidisciplinaire
  - chercheurs, documentalistes, informaticiens, …
- L'archivage est indispensable pour une meilleure valorisation et gestion des données
- Forte relation avec la science ouverte donc il est possible de mutualiser les efforts
- Faible sensibilisation à tous les niveaux (chercheurs, décideurs, …)
- L'archivage n'est pas une démarche seulement informatique
- Modèle économique au sens large (ressources budgétaires) nécessaire pour la pérennisation du service d'archivage
- Facteur humain très fort --- pas de cas typique -- caractère unique de chaque projet de recherche -- non automatisable

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

# Archivage des Données à l'IN2P3

Yonny CARDENAS

Journée archivage numérique des données de recherche

Grenoble, 20 Novembre 2019