



**PREDON**

# **PRÉSERVATION DES DONNÉES SCIENTIFIQUES**

Cristinel DIACONU

Centre de Physique des Particules de Marseille

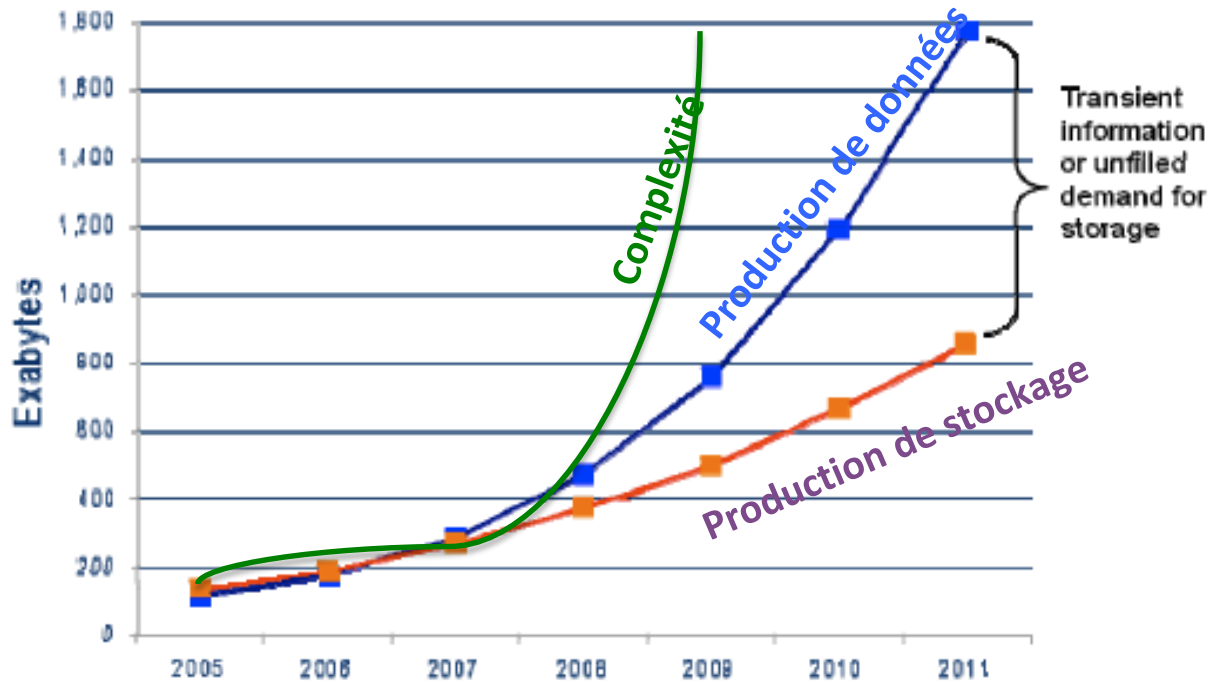
[diaconu@cppm.in2p3.fr](mailto:diaconu@cppm.in2p3.fr)

Pour le groupe PREDON



# Les données digitales sont fragiles

La capacité de stockage est physiquement dépassée depuis longtemps  
Complexité, hétérogénéité, origine, reproductibilité etc.



## Big Data: Les 6 'V'?

- Valeur
- Veridicité
- Vitesse
- Variété
- Volume
- Vulnérabilité???

Perte de données?

FIGURE 1.3: Information and Storage

Source: J. Gantz January 2008 (revised). Used with permission.

# Données scientifiques

## De plus en plus complexes

- Information riche, collectée par des capteurs versatiles

## Encore plus vulnérables:

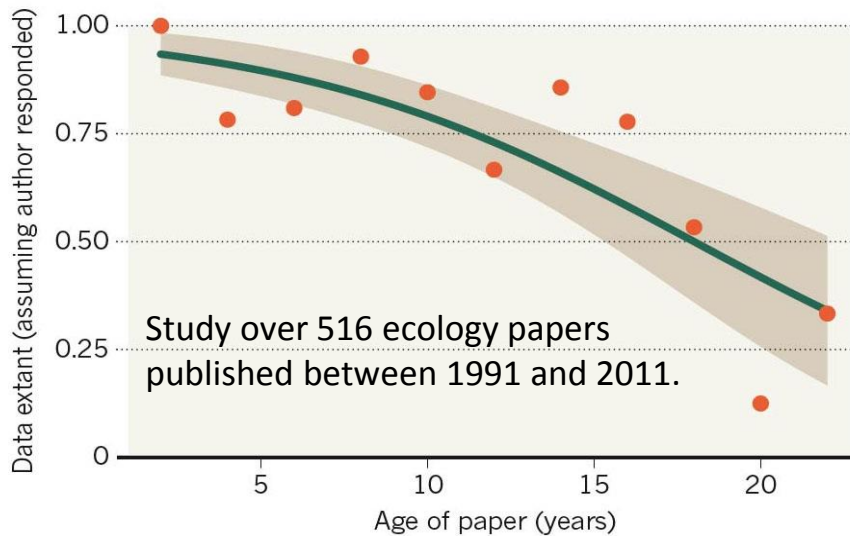
- modèle économique de la préservation à long terme quasi inexistant

## Motivation scientifique évidente

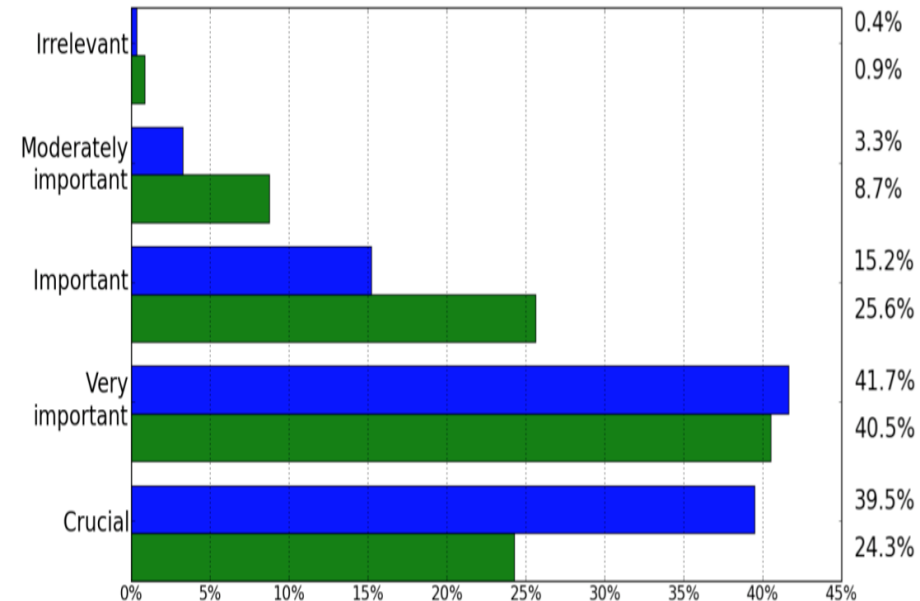
- Recherche à bas coût, retour sur l'investissement

### MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.

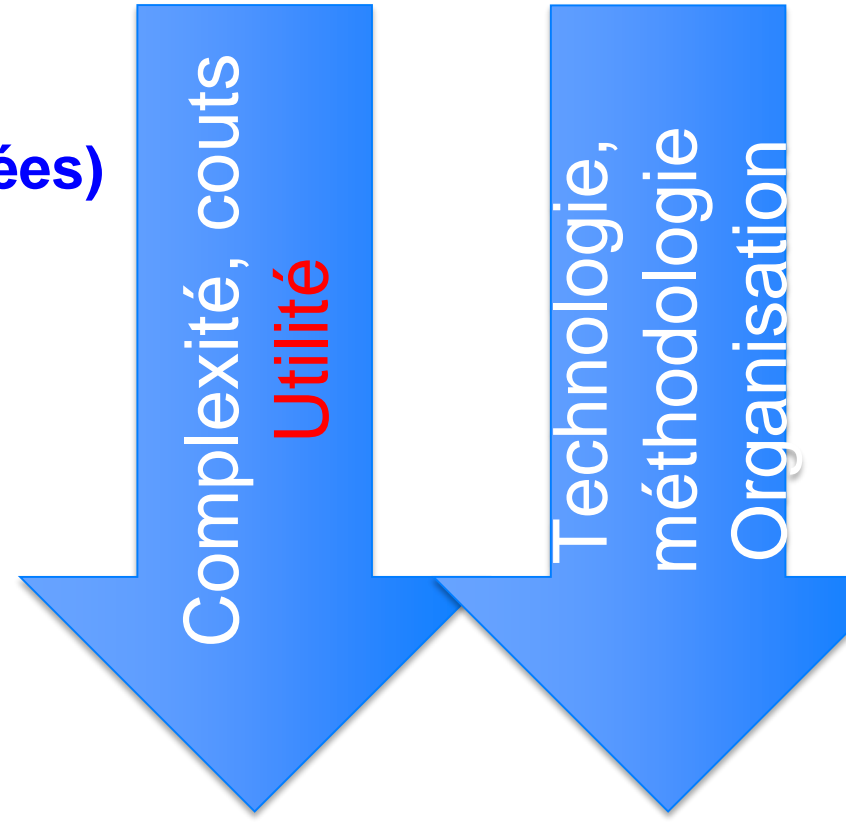
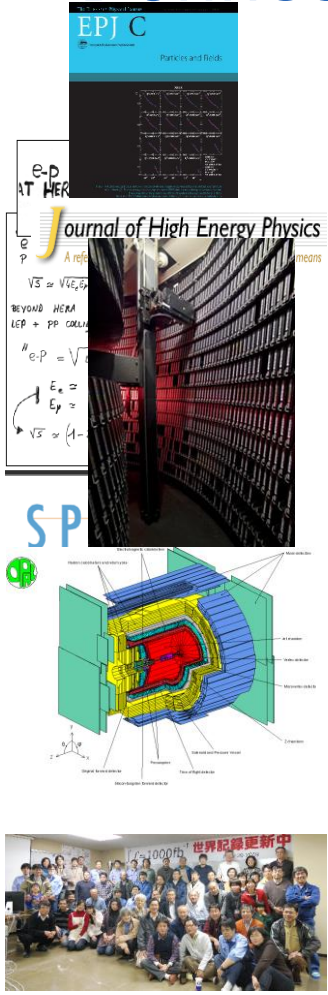


In your opinion, how important is the issue of data preservation ?  
(top/blue: theorists, bottom/green: experimentalists)



# Données Scientifiques

- Publications
- Documentation
- Donées (brutes+processées)**
- Meta-données
- Workflows
- Software
- Diffuse knowledge
- ....more...



Quel modèle de préservation pour les données scientifiques?

# Préambule

**Les données scientifiques ont un potentiel qui dépasse le cadre de recherche initial et qui doit être exploité à long terme**

- ◉ **Preservation  $\Leftrightarrow$  Accès ouvert**

**La préservation de données scientifique est économiquement avantageuse:**

- ◉ **Recherche à bas cout**

**Une technologies de frontière est nécessaire**

- ◉ **Préservation de toute la chaine « grise »**
- ◉ **Virtualisation, cloud computing, workflows....**

**La collaboration multi-disciplinaire est essentielle**

- ◉ **au niveaux national et international**
- ◉ **Projet PREDON: animation, R&D, architecture**



# Challenges

## Préserver les « octets »

- ◉ Supports, centres de données
- ◉ Coûts?
  - 2x taille initiale (1+1/2+1/4+....)

## Préserver les procédures

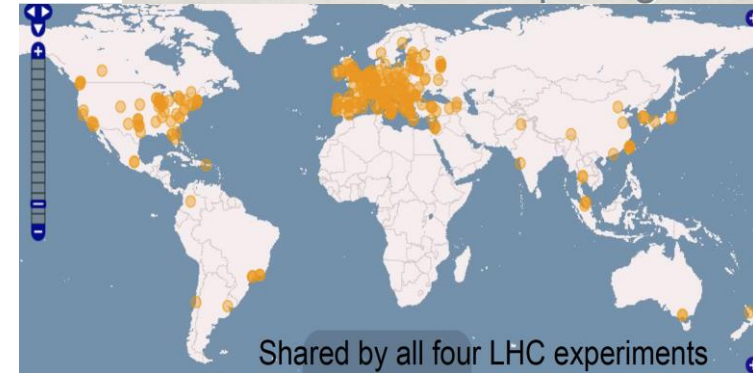
- ◉ Algorithmes, workflows etc.
- ◉ Software: complexe, fragile

## Préserver les connaissances

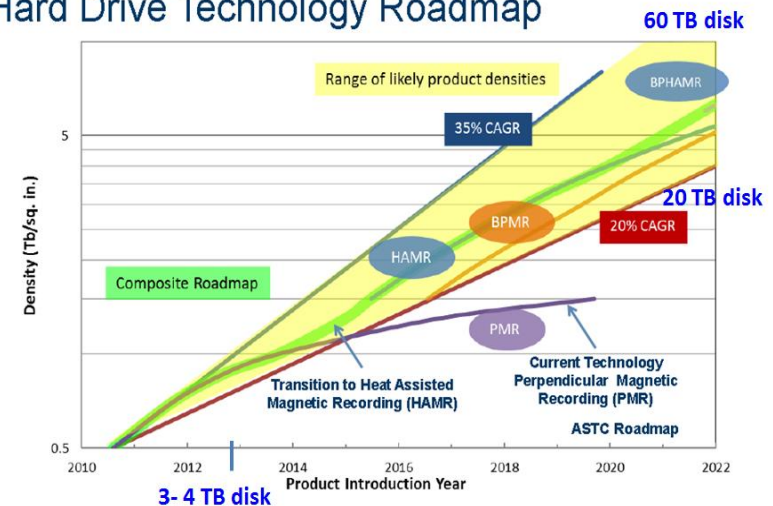
- ◉ Indexation, metadata, standards,...(OAIS)
- ◉ Documentation, connaissances
- ◉ Collaborations à long terme

Storing the data is not a problem: hard drives are cheap and getting cheaper. The challenge is preserving knowledge that is less commonly stored — the software, algorithms and reference

## Worldwide LHC Computing Grid



## Hard Drive Technology Roadmap



# Generic models for Data Preservation

## Technology preservation

- ◉ Freeze the hardware : limited capability, one day it will fall apart however



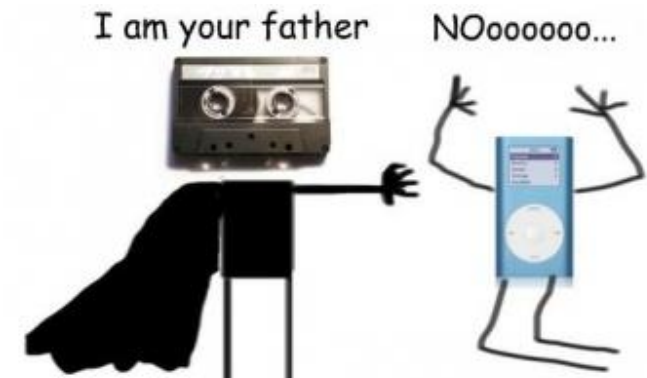
## Technology emulation

- ◉ Based on virtualisation
- ◉ Prepare it once (?), migrate the “middleware”



## Continuous migration

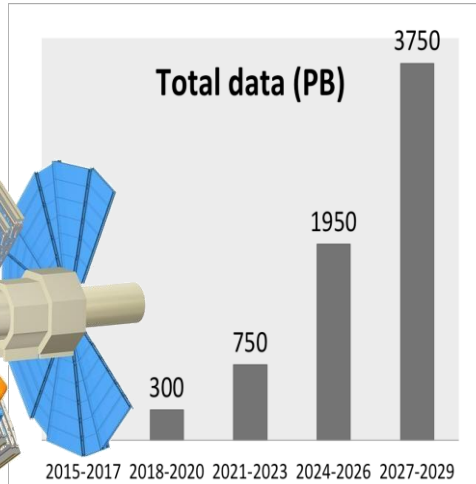
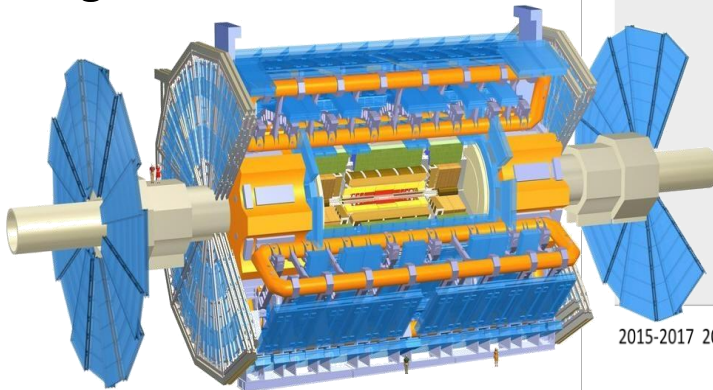
- ◉ Follow technology changes (adjust, redesign, recompile etc.....)
- ◉ Validation plays a central role



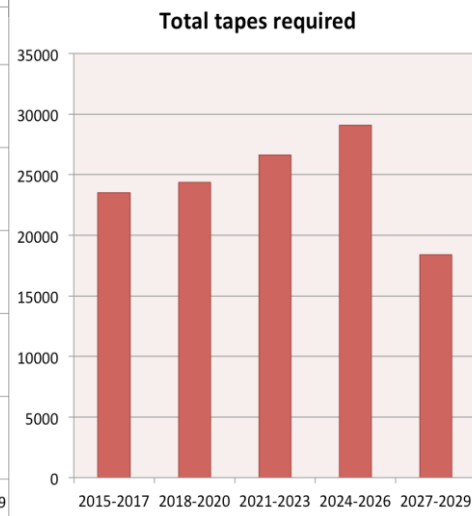
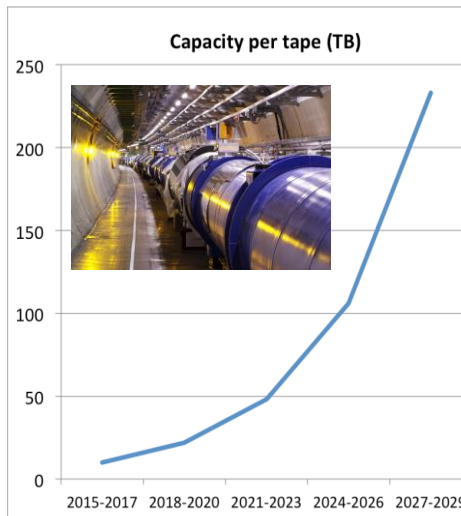
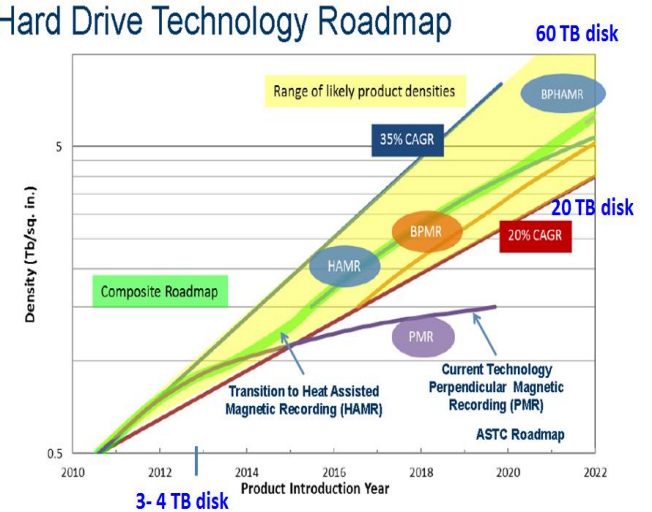


# The problem is not (only) the storage

Example:  
Large Hadron Collider

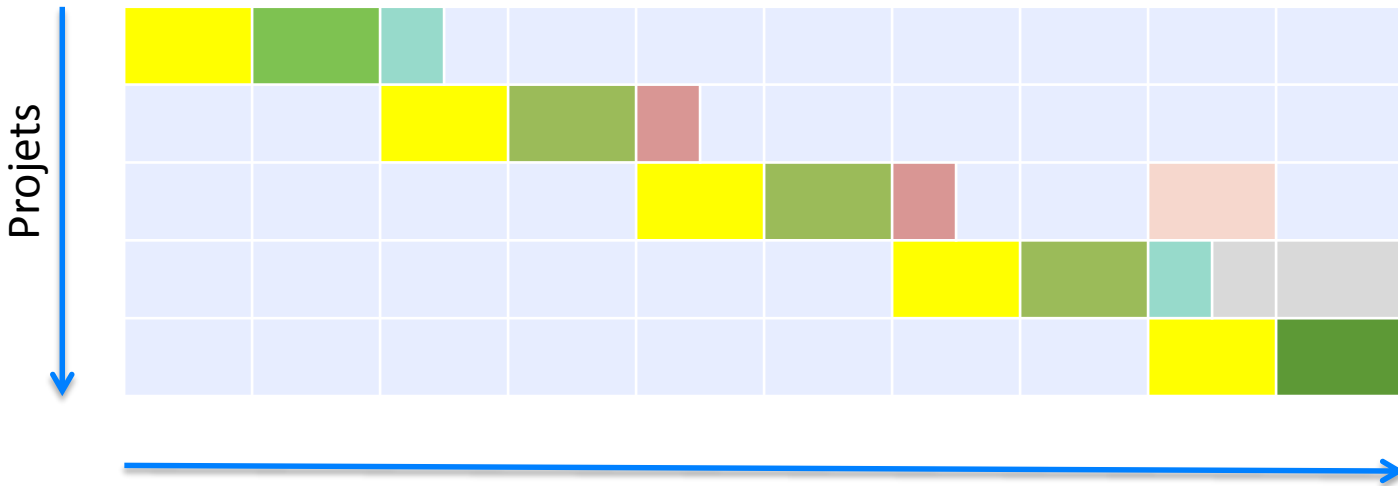
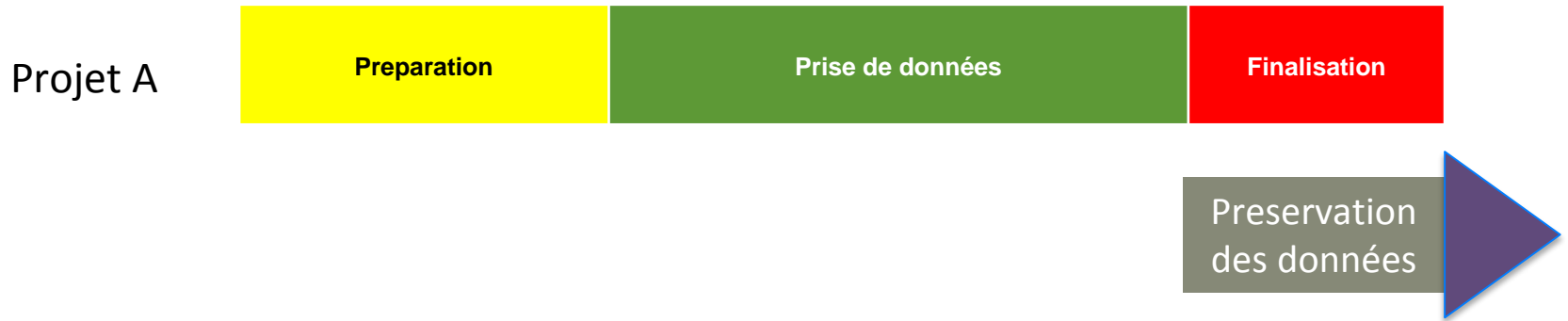


## Hard Drive Technology Roadmap





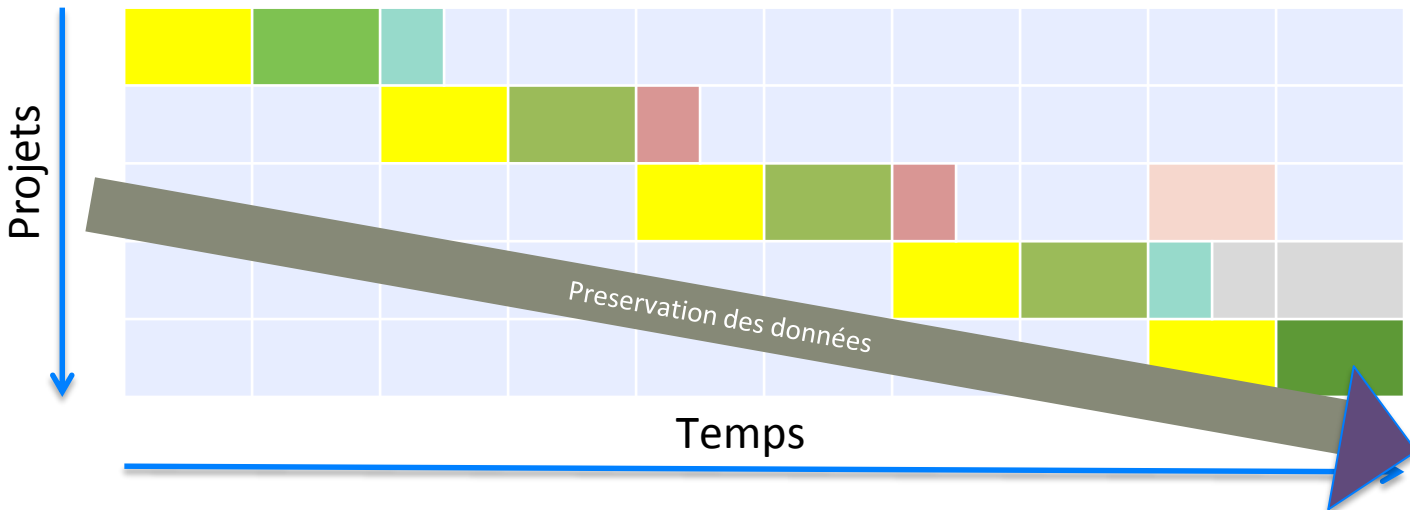
# Quand faut-il commencer à préserver?



# Quand faut-il commencer à préserver?



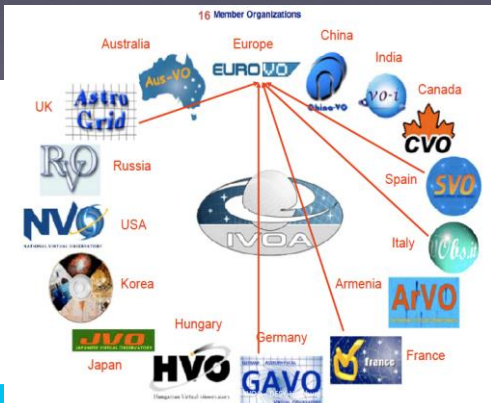
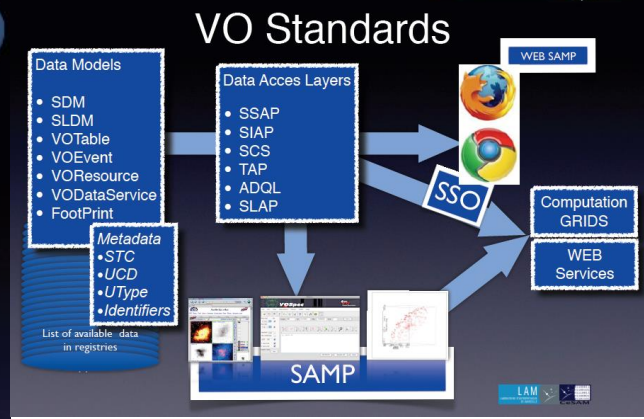
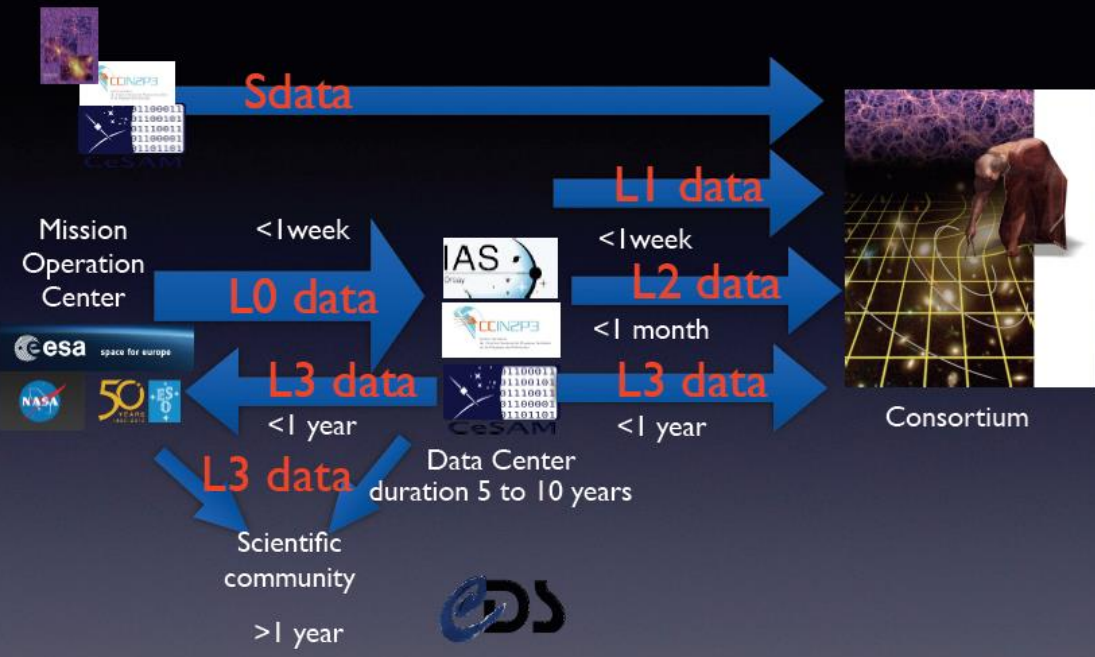
« Data Management Plan » doit inclure la préservation et l'accès à long terme



Programme cohérent de la préservation de données

# Astrophysique: Observatoires Virtuels

## Data Flux



<http://www.ivoa.org>

C. Diaconu

# Structuration dans la physique des particules

## DPHEP « Memorandum of understanding » signé par des agences de financement:

- Suisse(CERN), France, Japon, Finlande, Allemagne, Chine

## CERN: portal « open data » pour les données du LHC

C.Diaconu

## Collaboration Agreement for the DPHEP Project

BETWEEN:

The Partners of the DPHEP Project (the "Partners") set out in Annex 1 to the Collaboration Agreement,

CONSIDERING THAT:

(1) Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique;

(2) The Data Preservation and Long Term Analysis in High Energy Physics (DPHEP) project (the "Project"), an inter-experimental study group on HEP data preservation and long-term analysis, was initially formed by large collider-based experiments to investigate the technical and organizational aspects of HEP data preservation and convened by a Chair and a Project Manager as a panel of the International Committee for Future Accelerators (ICFA); Two reports were released, providing an analysis of the research case for data preservation and a detailed description of the various projects at experiment, laboratory and international levels;

(3) In its report of May 2012 (see Annex 2), the study group provided a concrete proposal for an international collaboration in charge of the Project and data management and policies in high-energy physics;

(4) The Partners have expressed their interest to take part in and contribute to the Project in order to implement the recommendations provided in the report referred to in Annex 2 and wish to formalize their collaboration through the present Collaboration Agreement;

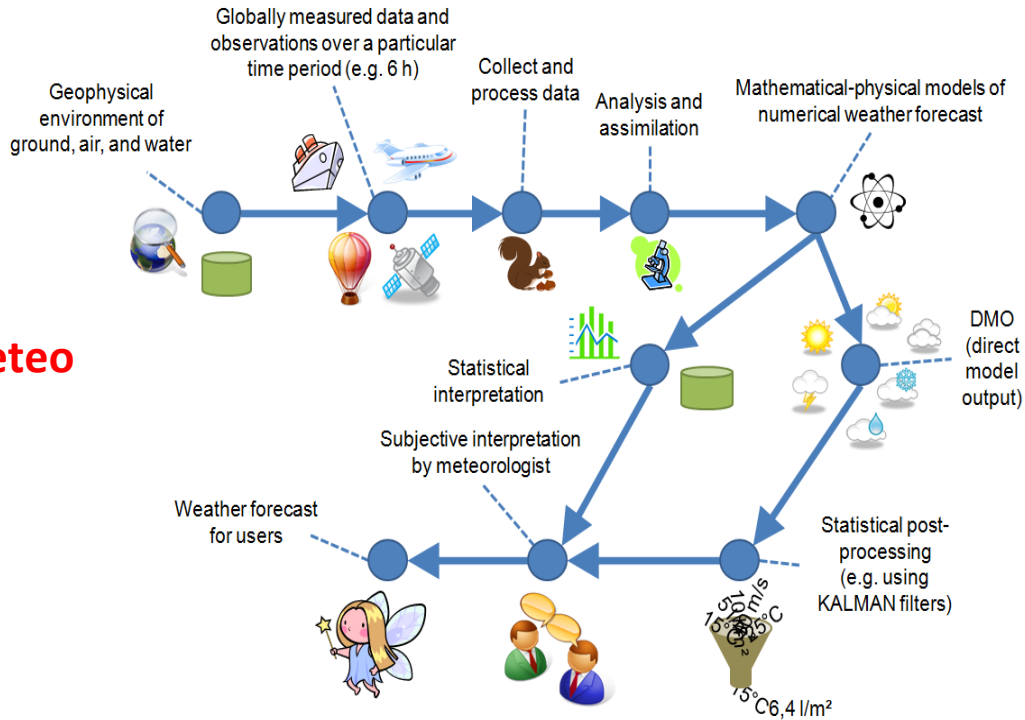
(5) The mutual benefit of the Partners that shall result from collaboration between them;



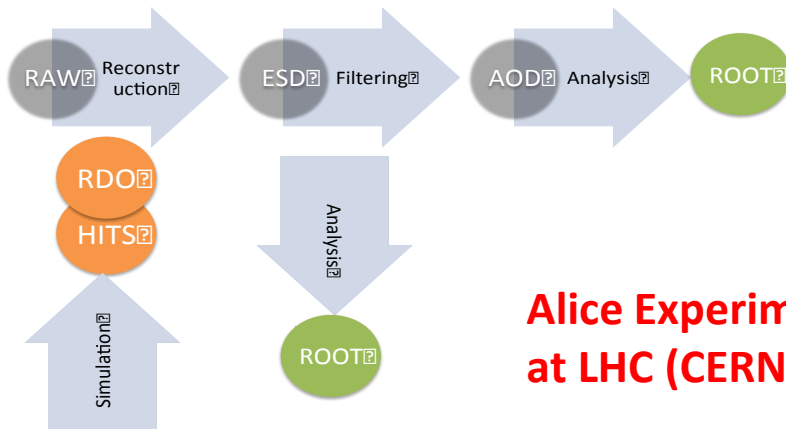
First DPHEP Collaboration Board

# Formats, workflows et préservation

**Meteo**



=



**Alice Experiment  
at LHC (CERN)**

Formats de données: standards?

Similarité entre les disciplines

Approche théorique rigoureuse  
Besoin et opportunité

Projet dans le cadre « Mastodons/Big Data » de la MI/CNRS et action dans Madics (GDR Big Data)

	Volume données	Complexité	Diversification des sources	Structuration au niveau international	Algorithmes et méthodologies pour la préservation
IN2P3 HEP	+++	+++	+	++	+
INSU, IRD Astrophysics Earth Sciences	++	++	++	+++	++
CINES INS2I IT, Algorithms, workflows	+	++	+++	+	+++

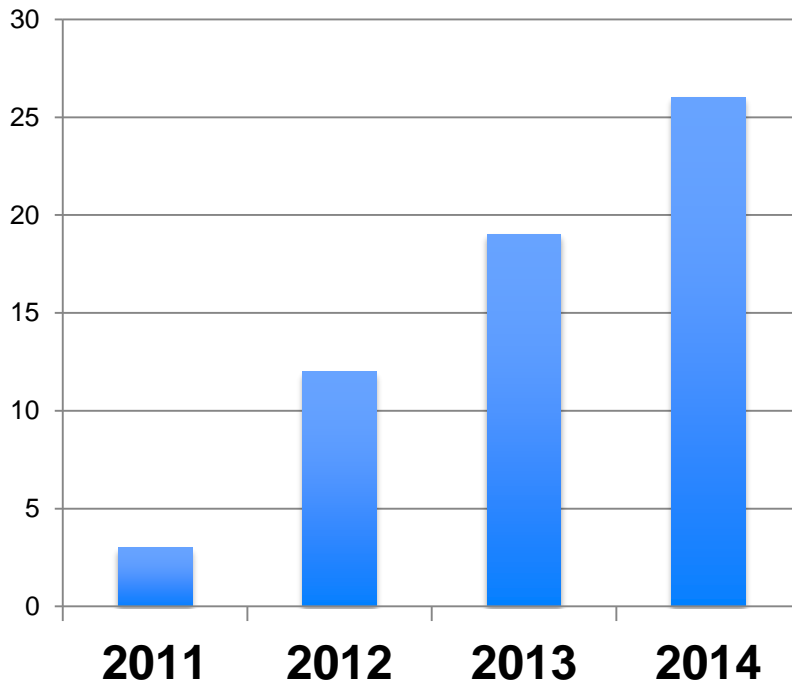
Harmonisation  
R&D

Architecture, Pilotage



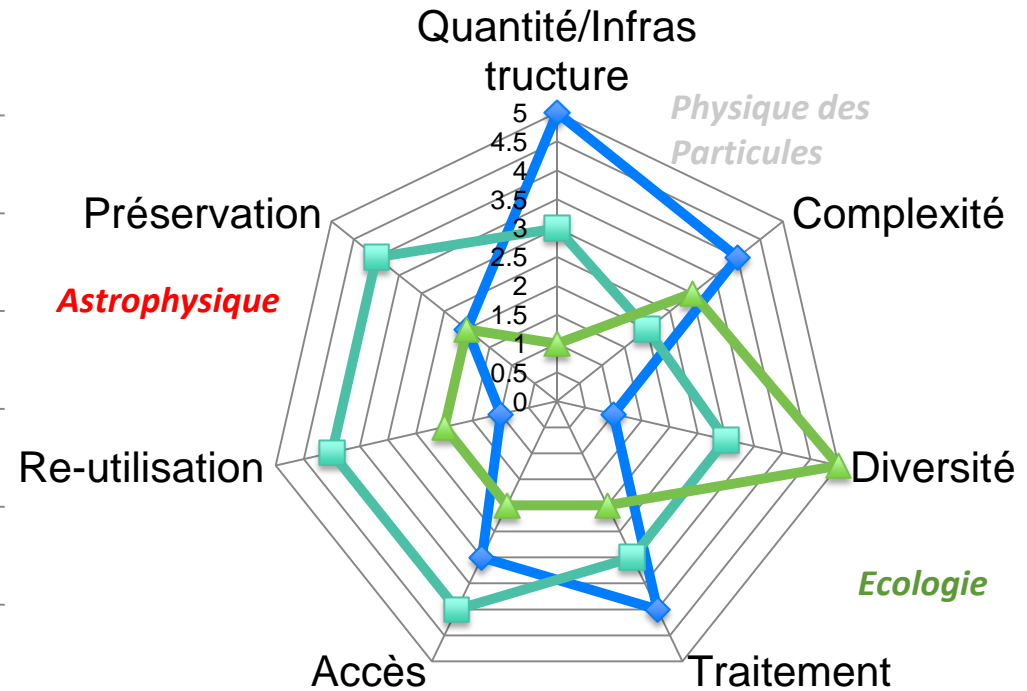
# Multidisciplinarité, complémentarité

## Personnes de contact PREDON



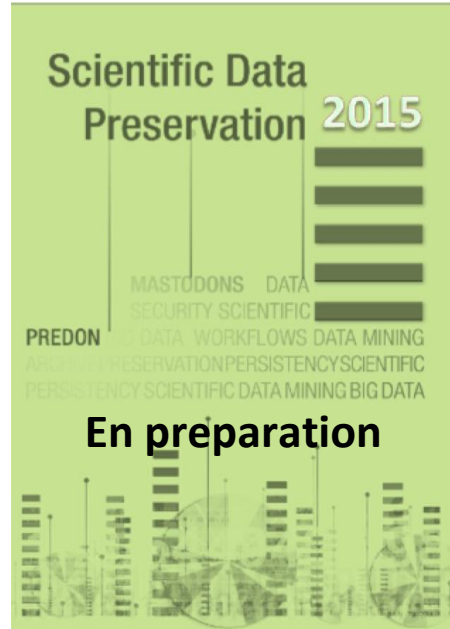
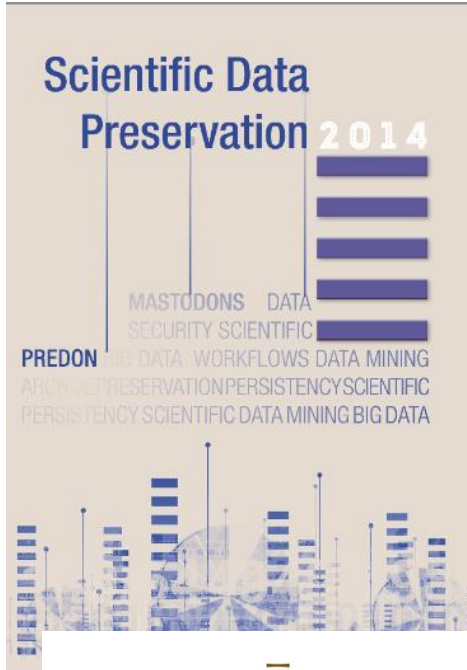
### Domaines:

physiques des particules, astrophysique, cristallographie, sciences de la terre, informatique, écologie (**IndexMed**), sciences de l'information, imagerie médicale, centres de calcul et stockage



Exemple sur 3 domaines

# Document PREDON 2014



## Scientific case

- 9 DATA PRESERVATION IN HIGH ENERGY PHYSICS  
C. Diaconu, S. Kraml
- 13 VIRTUAL DATA PRESERVATION  
C. Su
- 17 CRYSTAL DATA PRESERVATION  
D. Ch
- 21 SATELLITE DATA MANAGEMENT AND PRESERVATION  
T. Libourel, A. Laurent, Y. Lin
- 25 SEISMIC DATA PRESERVATION  
M. Schaming

## Scientific Case

...TION: A WORLD-WIDE INITIATIVE

## Methodologies

- 31 WORKFLOWS AND SCIENTIFIC BIG DATA PRESERVATION  
S. Benbernou, M. Lebbah
- 35 LONG-TERM DATA PRESERVATION  
D. Bo
- 39 CLOUD-BASED DATA PRESERVATION  
C. Cénn, M. Lebbah, H. Azzag
- 43 SCIENTIFIC DATA PRESERVATION, COPYRIGHT AND OPEN SCIENCE  
P. Mouron

## Methodologies

## Technologies

- 49 STORAGE AND DATA PRESERVATION  
J.-Y. M
- 51 REQUIREMENTS FOR SCIENTIFIC DATA AT CINES  
S. Cou
- 57 VIRTUAL ENVIRONNEMENTS FOR DATA PRESERVATION  
V. B

## Technologies

<http://informatique.in2p3.fr/li/?page=lettre&numero=27>



n°27 Avril 2014  
**La lettre IN2P3 Informatique**  
Réseau des Informaticiens de l'IN2P3 et de l'IRFU



Top départ pour la salle informatique « vallée » de Virtual Data !



Virtual Data est le groupe de travail du labex P2IO (réseau de tous les laboratoires du Sud de l'île de France impliqués dans la physique de l'infiniment petit, de l'infiniment grand et de l'étude des conditions d'apparition de la vie) dédié à la réflexion sur l'évolution des besoins informatiques des laboratoires du labex. Après des mois de réhabilitation d'un bâtiment de la vallée de l'université Paris-Sud, une longue phase de planification et de tests nombreux a permis de retrouver entre les salles informatiques historiques des laboratoires P2IO et la nouvelle salle « vallée » de VirtualData, un grand nombre de serveurs ont rejoint leurs nouveaux emplacements, concluant ainsi brillamment le déménagement d'une grande partie de l'infrastructure informatique de nos laboratoires de la vallée ! Ces machines utilisent maintenant presque totalement 18 racks sur les 30 prévus (pour une puissance totale de 400 kW) installés dans les 100 m<sup>2</sup> de la première tranche de « Virtual Data », un projet porté par le Labex, mené à bien, dans les temps et dans le budget prévu, par un groupe de travail d'informaticiens des différents laboratoires de P2IO avec l'aide du service infrastructure du LLN.

Interview  
"Assurer notre excellence dans le domaine du calcul pour servir nos communautés scientifiques reste une priorité"



Ursula Bassler, directrice adjointe scientifique de l'IN2P3 en charge de la physique des particules et du calcul

### Stockage



Le service iRODS à l'IN2P3

Depuis plus d'une décennie, l'IN2P3 comme beaucoup d'autres acteurs du monde scientifique, est le témoin d'une explosion des quantités de données produites par les projets scientifiques qu'il soutient. La gestion de ces masses de données est un enjeu de première importance : il faut à la fois assurer

### Développement



Pyrame, un framework de prototypage rapide pour systèmes online

Pyrame est un framework léger développé par le Laboratoire Laprice-Ringnet (IN2P3/École Polytechnique) qui permet de prototyper rapidement et facilement des systèmes online de complexité arbitraire. Pyrame est basé sur une série de modules, gérant chacun un matériel spécifique ou une fonctionnalité système. Ils communiquent entre eux au

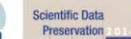
### Collaboratif



Accès nomade à Internet des personnels IN2P3 via eduroam.in2p3.fr

Le service eduroam vise à offrir un accès sans fil sécurisé à l'internet pour les personnels des établissements d'enseignement supérieur et de recherche lors de leurs déplacements en France et à l'étranger. Les utilisateurs d'un établissement membre du

### Préservation de données



Scientific Data Preservation

PREDON : La

### Imprimer

Agenda  
EGI Community Forum, 19-23 mai - Helsinki  
Le prochain Community Forum d'EGI (European Grid Infrastructure) se tiendra du 19 au 23 mai (...)  
en savoir plus

HEPIX, 19-23 mai - LAPP  
The Local Organizing Committee is pleased to announce that the registration and abstract (...) en savoir plus

# SPADON

## Les supports pour l'archivage « passif » à long terme

- Etudes de vieillissement et préconisation des supports à longue durée de vie
- Caractérisation chimique du vieillissement!
- Rapport PSN des académies



C.Diaconu

## Recommandations du rapport PSN

Un message d'alerte au grand public, aux établissements et à l'administration, doit être lancé. Le problème n'est pas spécifique à la France, mais mondial. Une action au niveau européen, ou dans le cadre de l'UNESCO, semblerait souhaitable, afin de faire prendre conscience de l'urgence d'une politique concertée dans ce domaine. A l'échelon national et européen, nous proposons quatre recommandations:

1. Débloquer les études sur le sujet.
2. Éviter la perte des compétences dans le privé et le public
3. Favoriser l'innovation et l'apparition d'une offre industrielle de qualité
4. Élaborer une véritable politique d'archivage numérique



# Etudes et impact de la non reproductibilité en bioinformatique

S. Cohen-Boulakia  
Ch. Blanchet  
O. Collin

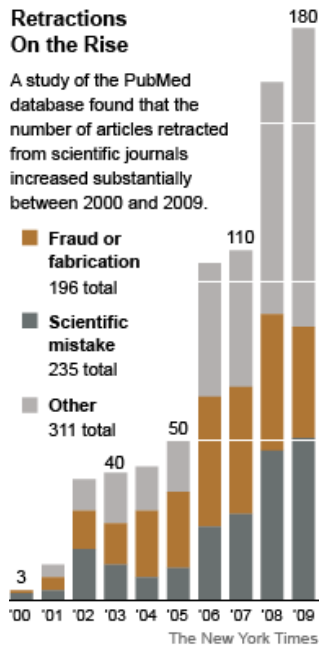
## Nekrutenko & Taylor, Nature Genetics (2012)

- 50 articles résultats obtenus avec un outil → 7/50 (14%) reproductibles

## Alsheikh-Ali et al, PLoS one (2011)

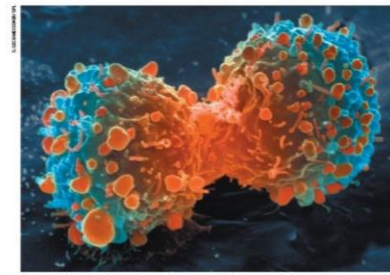
- 10 articles dans le top-50 journaux IF → 47 /500 (9%) reproductibles

## Conséquences diverses



Articles retractés (NewYork Times)

Conséquences tragiques (pré-essais cliniques)



### Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex of disease, although we in the cancer field hoped that this would lead to more effective drugs. Ironically, our ability to translate this knowledge into clinical practice has been limited. The high cost and the complexity of drug development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will be tested in patients. Investigators must reassess their approach to translating discovery research into the clinic and measure its impact. Many factors are responsible for the high failure rate, and understanding the full range of difficulties of this disease, including the limitations of preclinical testing, is essential.

47/53 "landmark" publications could not be replicated [Begley, Ellis Nature, 483, 2012]

### Must try harder

Too many sloppy mistakes are creeping into scientific papers, at the data – and at themselves.

### Error prone

Biologists must realize the pitfalls massive amounts of data.

### If a job is worth doing, it is worth doing twice

Researchers and funding agencies need to put a premium on ensuring that results are reproducible, argues Jonathan F. Russell.

The case for open computer programs

### Six red flags for suspect work

C. Glenn Begley explains how to recognize the preclinical papers in which the data won't stand up.

Know when your numbers are significant

# IndexMED

## Objectif principal :

Développer la culture des bases de données et leur utilisation efficace dans le milieu de la recherche en écologie et biodiversité.

Acquisition automatisée

Systèmes de reconnaissance automatique

Outils d'aide à la détermination

Data management

Indexation de la donnée

Gestion de la qualité et réutilisation de la donnée

Qualification et ontologie

Fouille de données

Conservation des données (Nouveau : les variables essentielles de biodiversité)

Représentations des données

...



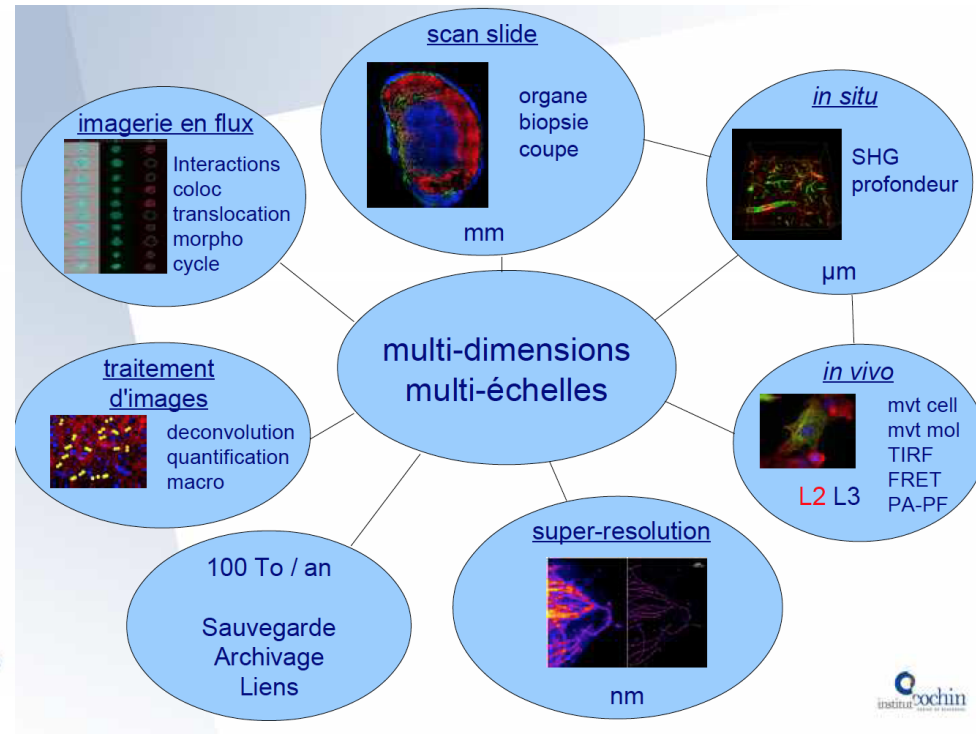
## Organisation d'un workshop IndexMED-PREDON en mars 2015

# Imagerie Cellulaire et le Big Data

P. Bourdoncle

## Plate-Forme Cochin Imagerie

- 13 systèmes d'acquisition
- 300 utilisateurs par an
- 1 serveur de transfère
- 4 stations d'analyse et de traitement d'image



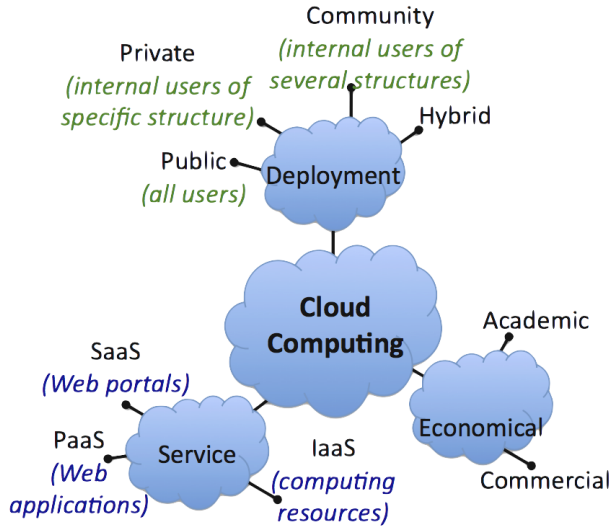
## Un potentiel important de coopération:

- Structuration, méthodes, technologies



# Préservation de données dans le « cloud »?

C. Cavet « Cloud technology for algorithm preservation » **PREDON workshop APC 4-6 Nov, 2014**



StratusLab

Home | Endorsers | Query | Upload | About

## Metadata

Show 10 entries

Search:

Status:

valid

Location:

all

Filter:

Search os

Search version

Search arch

Search endorser

Search kind

Sort by:

### CentOS v6.2 x86\_64

Endorser: [cecile.cavet@apc.univ-paris7.fr](mailto:cecile.cavet@apc.univ-paris7.fr)  
Identifier: [EvhQ9Mw\\_DUEI7Ykfs20vY0gWsZK](#)  
Created: 2013-10-15T09:39:02Z  
Kind: machine



Base image. Allows both standard StratusLab and cloud-init contextualization mechanisms. Image only has root account configured. Only logins via ssh keys are allowed. The root disk has 24 GB of space.

[More...](#)

### Ubuntu v12.04 x86\_64

Endorser: [images@stratuslab.eu](mailto:images@stratuslab.eu)  
Identifier: [KBhcU87Wm5IZNOXZYGHrczGekwp](#)  
Created: 2013-10-01T07:45:13Z  
Kind: machine



Ubuntu 12.04 base image automatically created by StratusLab. Configured only with a root user. The firewall in the image is disabled, IPv6 is enabled, and SELinux disabled. The root disk has 12GB of space. This image allows both standard StratusLab and cloud-init contextualization mechanisms. A swap volume is expected to be provided on /dev/sdb.

[More...](#)

### dummys v0.0 i686

Endorser: [loomis@lal.in2p3.fr](mailto:loomis@lal.in2p3.fr)  
Identifier: [GWE\\_nifKGCcXIFk42XaLrS8LQFJ](#)  
Created: 2013-09-03T12:32:06Z  
Kind: machine

[More...](#)

## Exemple: StratusLab

(<http://stratuslab.eu/index.html>)

End-user client

**MarketPlace (OS collection)**

Persistent disk Web interface

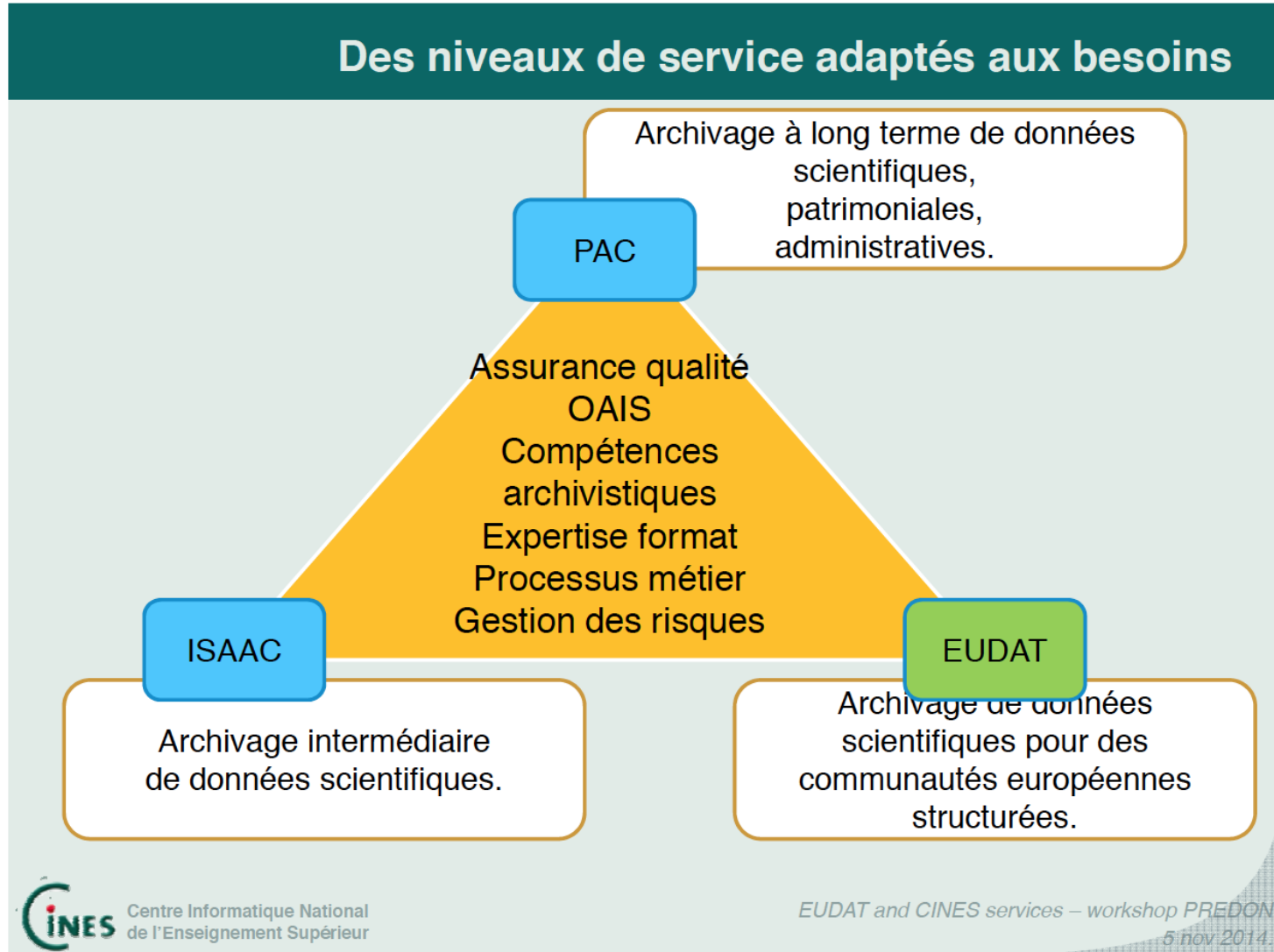
Ressource monitoring

Disk images have **6 months of validity**

OS update/upgrade for security.

Virtualisation/cloud need to be tuned for long term

## Projets CTNFS

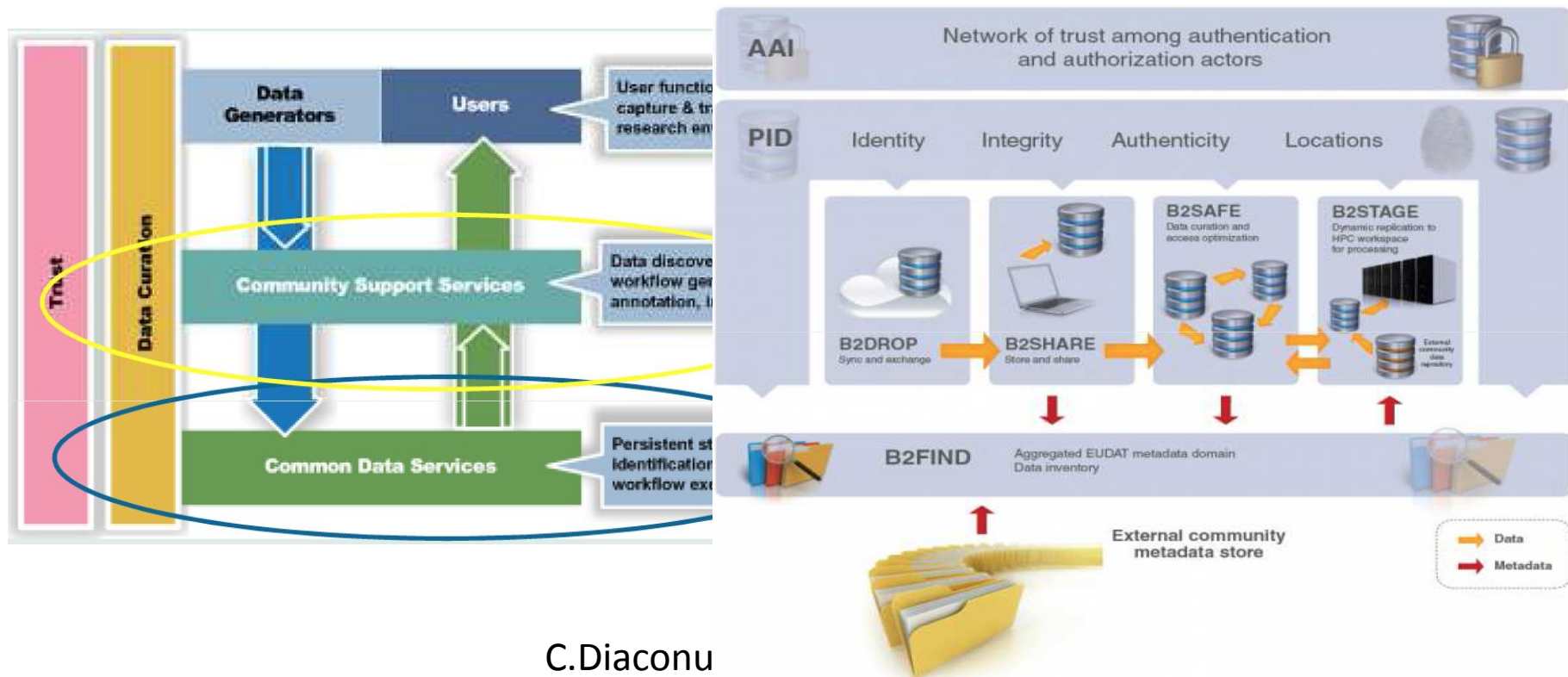


# EUDAT

**EUDAT will focus on building this generic data infrastructure layer and offer a trusted domain for long term data preservation accompanied with related services to store, identify, authenticate and mine these data.**

**Close collaboration with the Communities.**

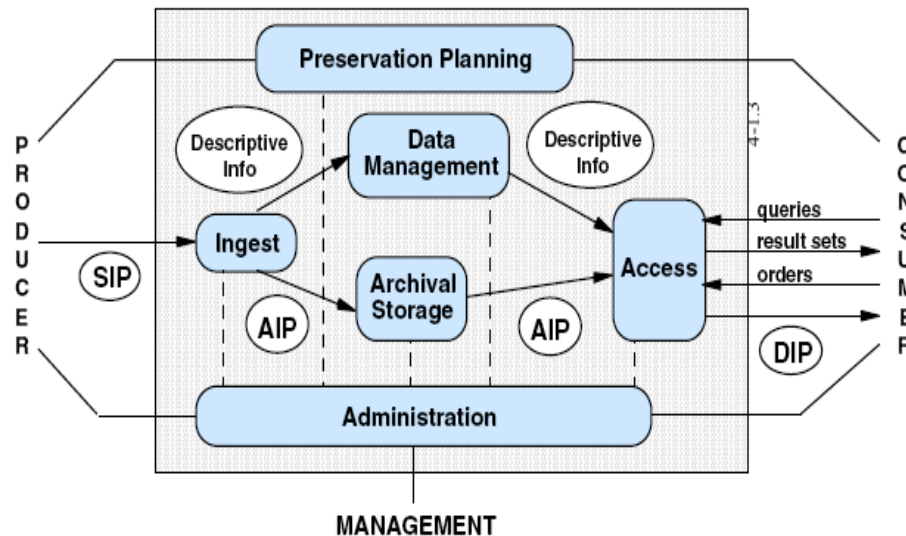
- Core services must match the requirements of the communities.
- Community services can also be incorporated into the common data service infrastructure when they are of use to other communities.



# Open Archive Information System OAIS

**OAIS = Modèle conceptuel et fonctionnel destiné à la gestion, l'archivage et à la préservation à long terme de documents numériques.**

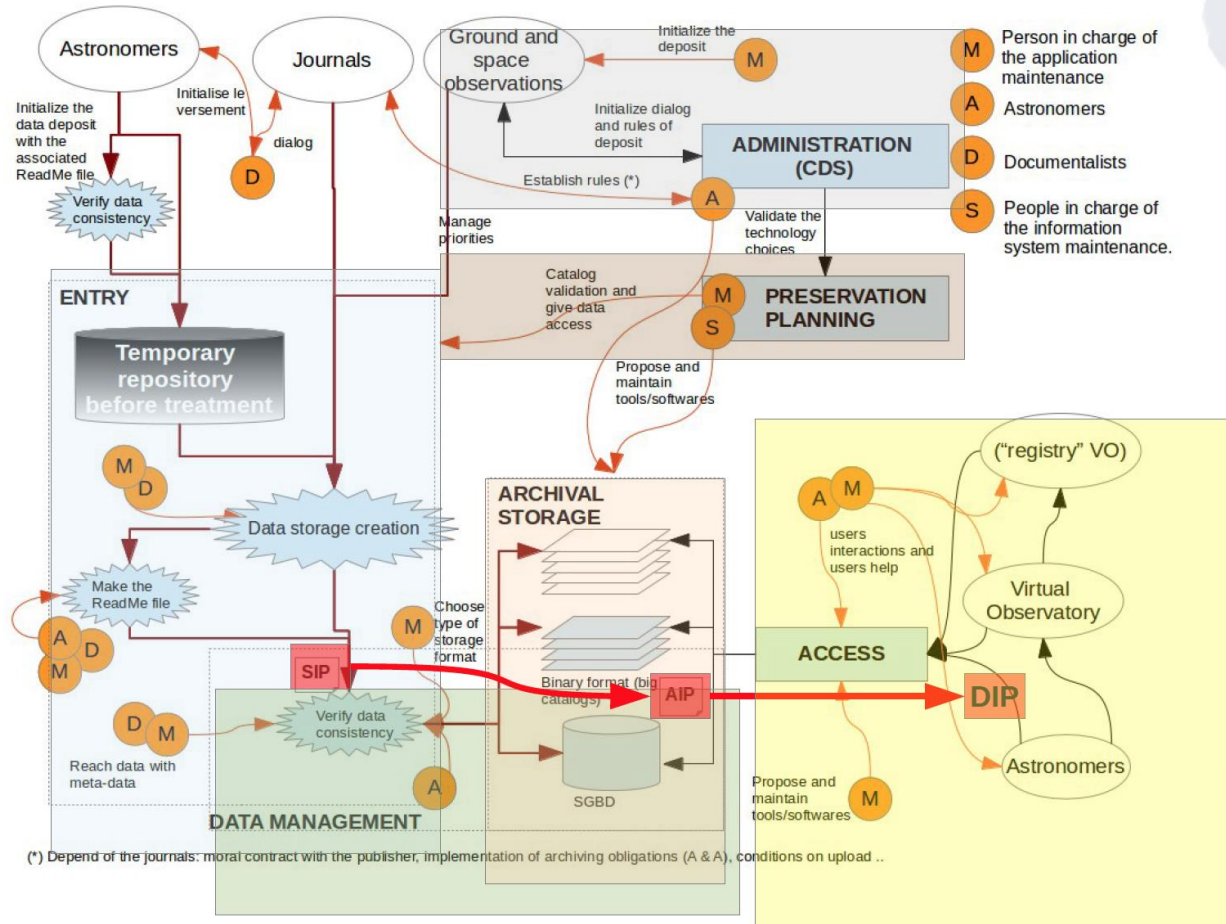
- Définit les acteurs/responsabilités dans le SI :
  - producteur/utilisateur/manager
- Définit les flux d'informations
  - → les paquets OAIS : SIP (en entrée), AIP (archives), DIP (diffusion)
- Définit des normes/recommandations « ouvertes »
  - exemple : modalités de versements des données



C.Diaconu

# Centre de données de Strasbourg

## Architecture OAIS-like pour la base de données Vizier



1,000,000 requetes/jour sur les services du CDS.





# Workshop PREDONx

## 9 Décembre 2015, Observatoire de Strasbourg

### PREDONx Workshop on Scientific Data Preservation

9 December 2015  
Centre de Données astronomiques de Strasbourg  
Europe/Paris timezone

Présentation

Programme scientifique

Agenda

Inscription

Hébergement

Lieu de la réunion

Programme sociale

PREDON

Video/Audio Conférence

Poster PREDON

Liste des participants

#### **PREDONx 2015 : Atelier sur la Préservation des Données Scientifiques**

**Mercredi 9 décembre 2015,**

**Centre de Données Astronomiques de Strasbourg**

Avec l'entrée dans l'ère digitale, l'humanité est devenue un gigantesque producteur de données scientifiques, complexes et aussi uniques pour la plupart. L'investissement humain et financier pour obtenir ces données est significatif et leur préservation à long terme plus que nécessaire.

Afin d'exploiter d'une manière intelligente l'effort investi pour des projets de recherche et d'observation, les programmes scientifiques doivent intégrer de manière cohérente la politique de sauvegarde et d'accès aux données à long terme.

L'atelier PREDONx 2015 est organisé par le projet PREDON, développé au sein du programme de grandes masses de données MASTODONS de la Mission Interdisciplinarité du CNRS. Le groupe opère aussi en tant qu'action au sein du GDR Madics depuis 2015.





09:00 - 10:30

## Case for Preservation

*Scientists spend a significant part of their time to design, collect and analyze data. While the lifetime of a project is often identified with the lifetime of its data, a clear and strong case have been presented in a majority of disciplines to preserve and re-use the scientific data, well after the initial project ends. In this session, concrete examples of scientific projects where data preservation is relevant will be discussed, in order to emphasize the need for a coherent long term perspective of scientific data preservation.*

09:10     **The golden mine of the future: scientific data preservation 20'**

Speaker: Cristinel Diaconu (CPPM, Aix-Marseille Université, CNRS/IN2P3 (FR))

09:35     **l'identification de provenance des données dans l'Observatoire Virtuel 20'**

Speaker: Mireille Louys (CDS/iCUBE)

10:00     **IVOA et les workflows scientifiques 20'**

Speaker: Schaaff Andree

10:30 - 11:00

## Pause Café

11:00 - 12:30

## Methodology for Data Preservation

*Methods, practices and projects for data preservation: community projects, work on data preservation standards, exchanges with libraries and information sciences, policies, legal aspects of data preservation.*

11:00     **Le projet pluri-disciplinaire IDV (Imagerie du vivant) de l'Université Sorbonne Paris Cité et quelques réflexions / méthodologies liées aux données. 25'**

Speakers: Christophe Cérin (urn:Google), Leila Abidi (Université Paris XIII)

11:30     **Nouvelles infrastructures numériques pour la recherche à USPC. 20'**

Speaker: Leila Abidi (Université Paris XIII)

11:50     **Open Data at the Large Hadron Collider 20'**

Speaker: TBC

14:00 - 15:45

## Technologies for Data Preservation

*Hardware providers, computing centers, industry actors etc. are providing various pieces and parts to preserve digital data long term; are these adapted for scientific data preservation? Do we understand the requirements and the limits of the present technology?*

14:00 **Petasky project (TBC) 20'**

Speaker: Christian SURACE (CNRS)

14:20 **Machine virtuelle pour garantir l'accès à long terme aux données et logiciels complexes 20'**

Speaker: Vincent Joguin (Eupalia SAS)

14:40 **Projet SEANOE 20'**

Speaker: Frédéric Merceur (IFREMER/Brest)

15:00 **Workflows for data preservation 20'**

Speaker: Salima Benbernou (Université Paris Descartes)

15:45 - 16:15

Pause Café

16:15 - 17:40

Discussion

19:00 - 21:30

Social Dinner

*Registered participants (who have confirmed their attendance at the dinner) are invited to join the social dinner (location to be determined).*

Location: To be announced