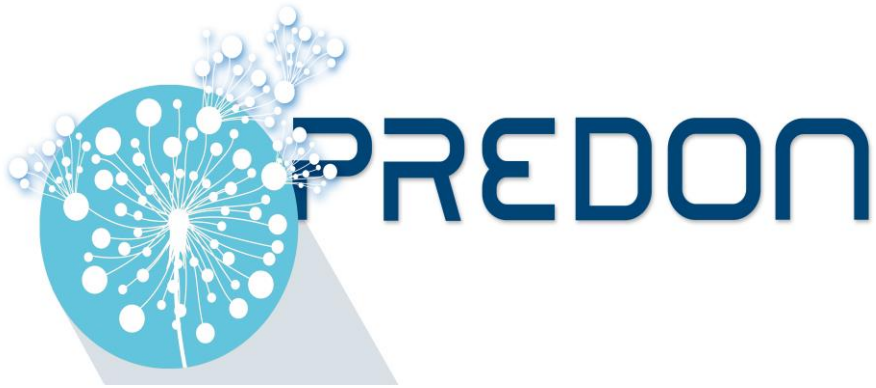


PRÉSERVATION DES DONNÉES SCIENTIFIQUES

Cristinel DIACONU
Centre de Physique des Particules de Marseille
diaconu@cppm.in2p3.fr

Pour le groupe PREDON

CENTRE DE PHYSIQUE DES PARTICULES DE MARSEILLE
CPPM

Aix*Marseille université

Les données digitales sont fragiles

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

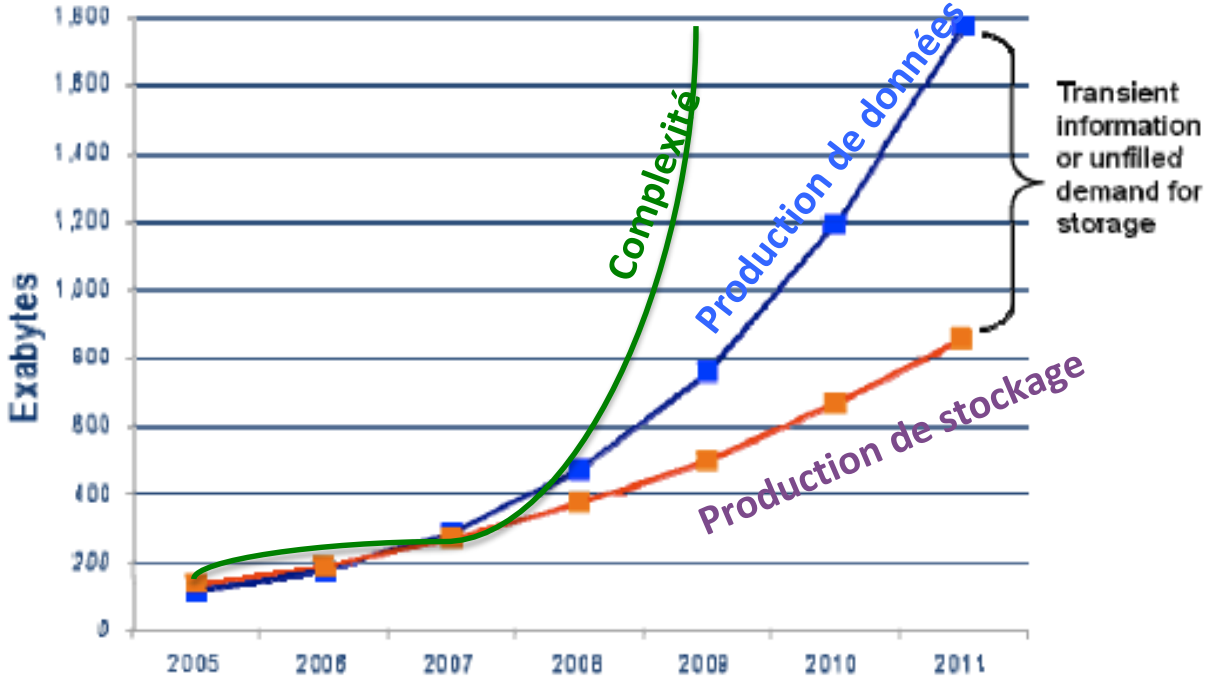


FIGURE 1.3: **Information and Storage**
Source: J. Gantz January 2008 (revised). Used with permission.

Perte de données?

Big Data: Les 6 'V'?

- Valeur
- Veridicité
- Vitesse
- Varieté
- Volume
- Vulnérabilité???

Données scientifiques

De plus en plus complexes

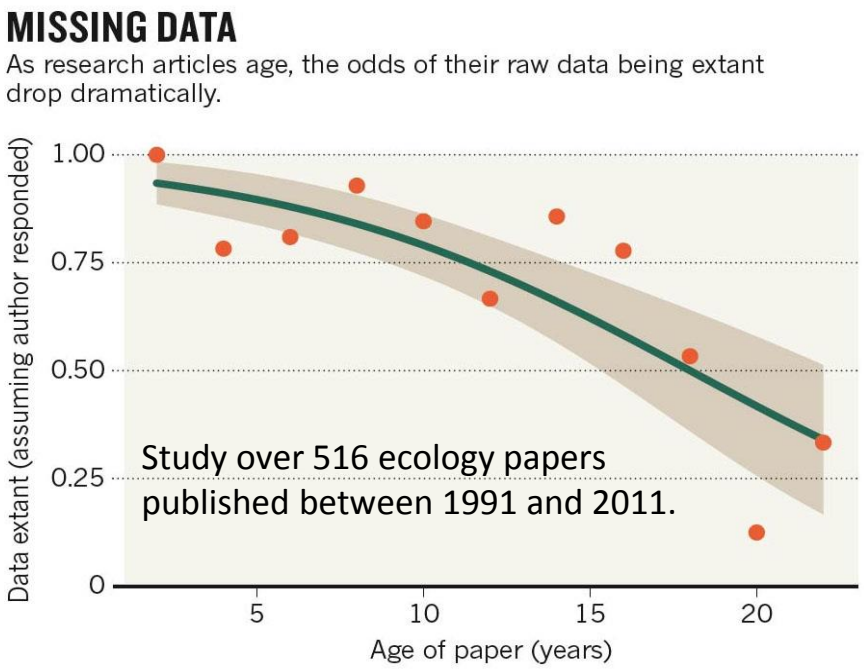
- Information riche, collectée par des capteurs versatiles

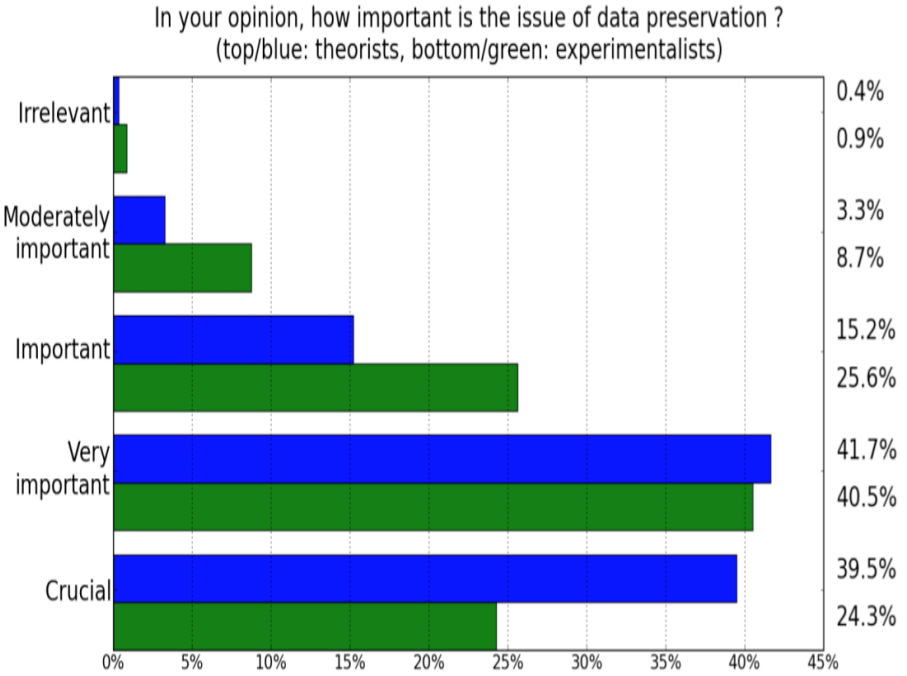
Encore plus vulnérables:

- modèle économique de la préservation à long terme quasi inexistant

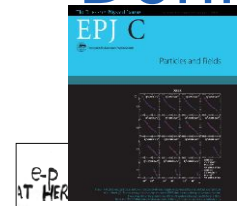
Motivation scientifique évidente

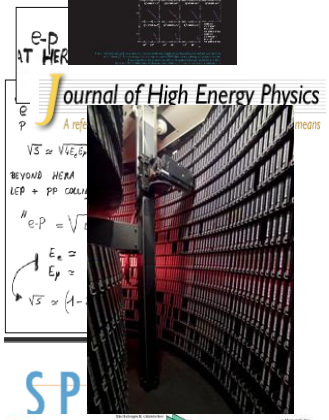
- Recherche à bas coût, retour sur l'investissement

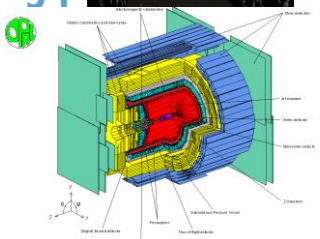




Données Scientifiques







Publications

Documentation

Donées (brutes+processées)

Meta-données

Workflows

Software

Diffuse knowledge

….more…

Complexité, couts
Utilité

Technologie, méthodologie Organisation

Quel modèle de préservation pour les données scientifiques?

C.Diaconu

Préambule

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

- **Preservation <=> Accés 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 »

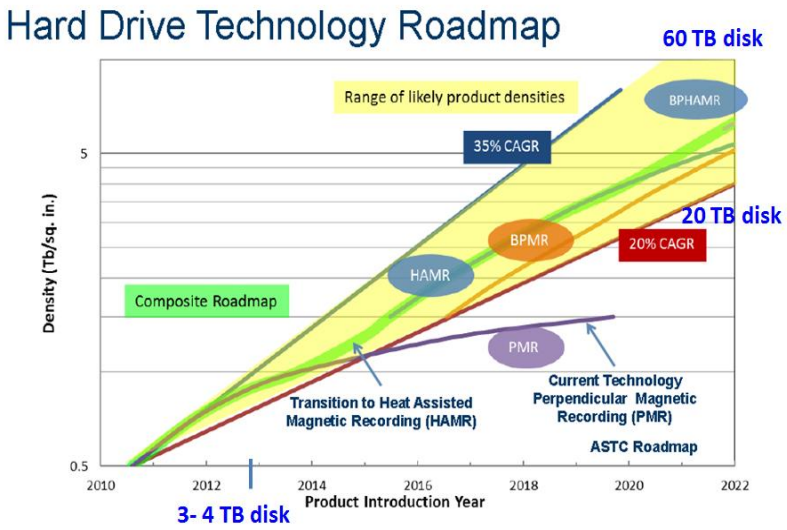
- Supports, centres de données
- Couts?
  - 2x taille initiale (1+1/2+1/4+….)

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

nature

Generic models for Data Preservation

Technology preservation

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

Technology emulation

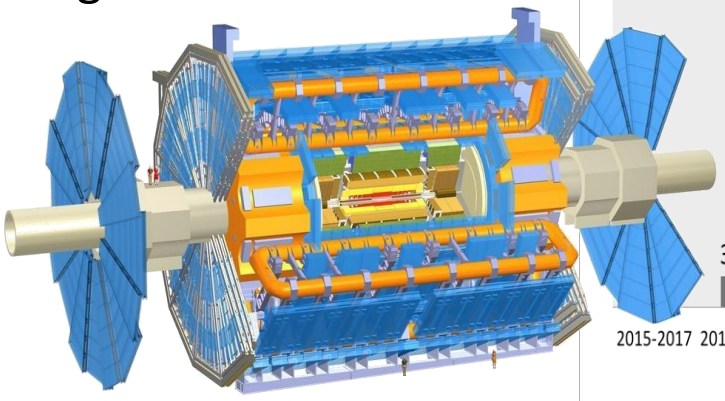
- **Based on virtualisation**
- **Prepare it once (?), migrate the “middleware”**

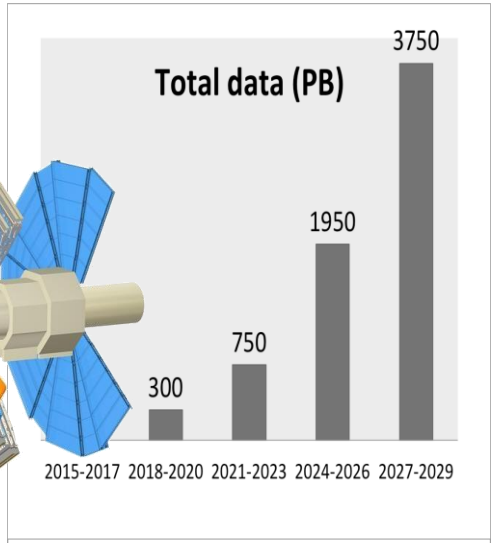
Continuous migration

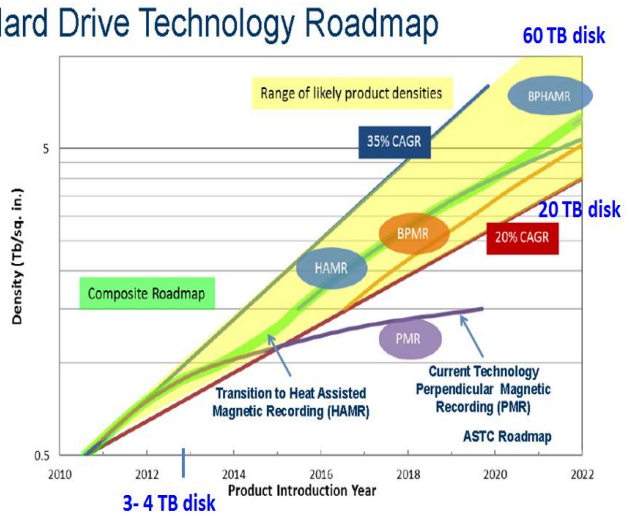
- **Follow technology changes (adjust, redesign, recompile etc….)**
- **Validation plays a central role**

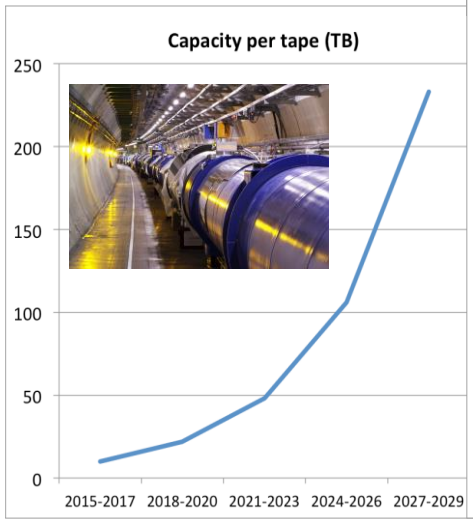
The problem is not (only) the storage

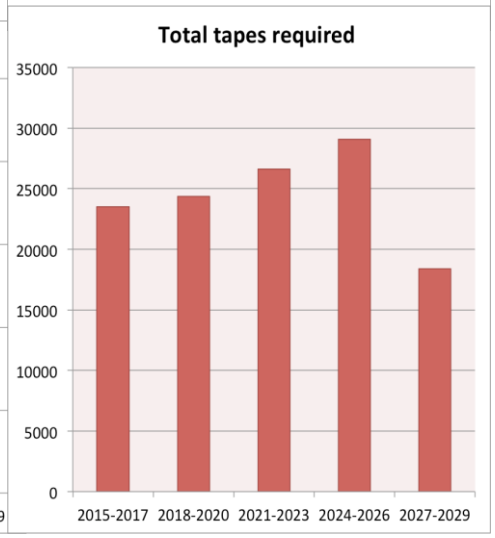
Example:
Large Hadron Collider



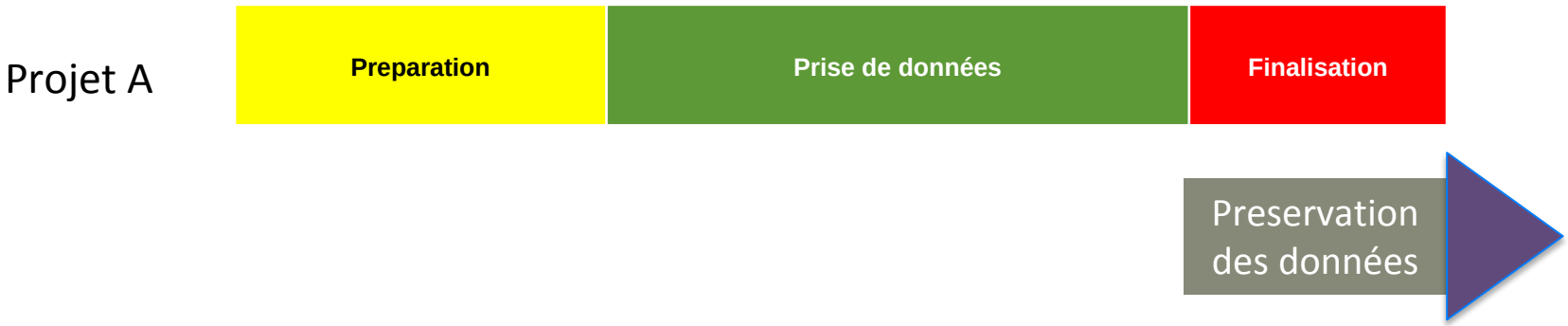


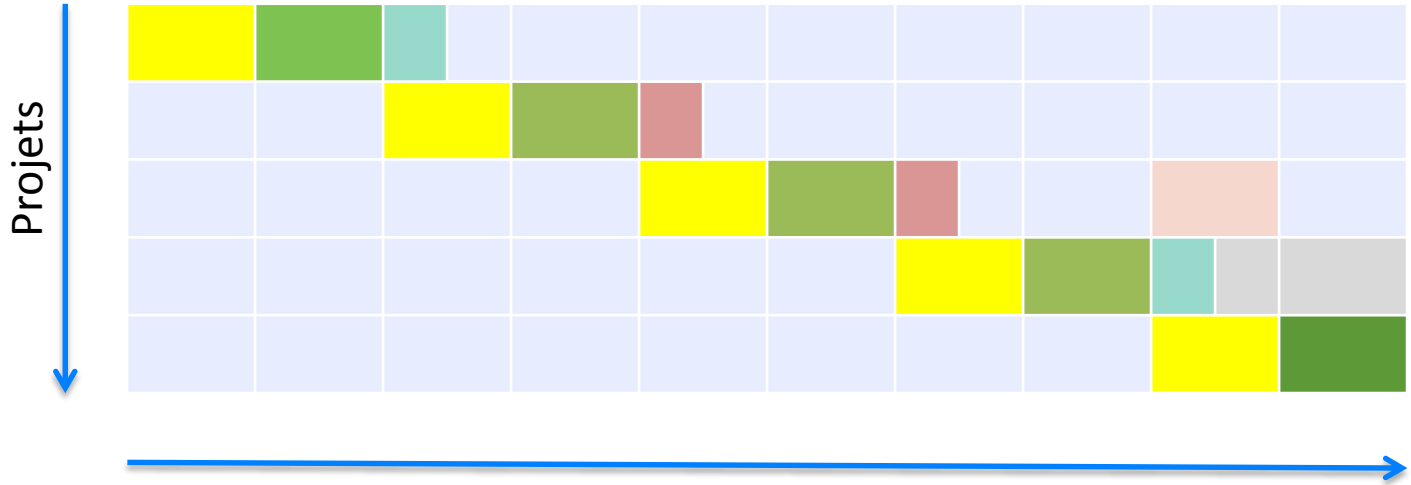






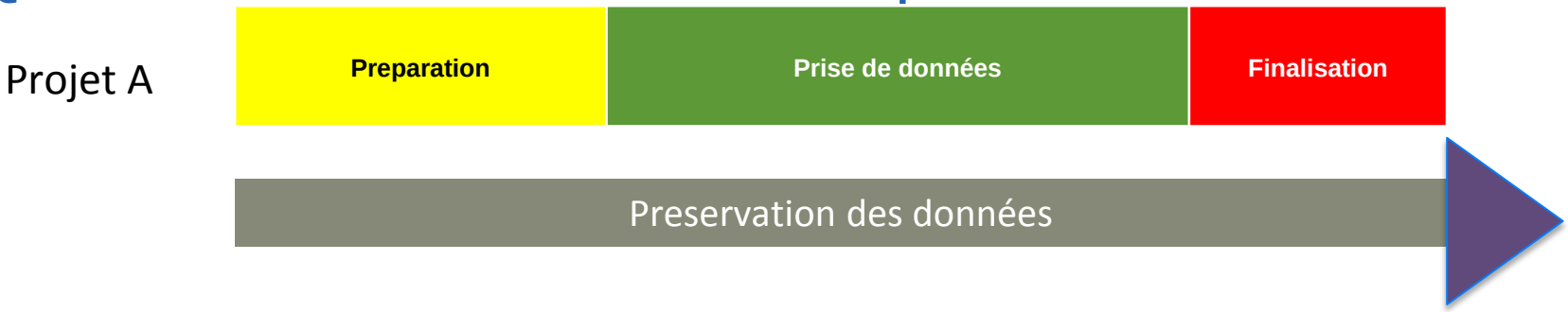
Quand faut-il commencer à préserver?



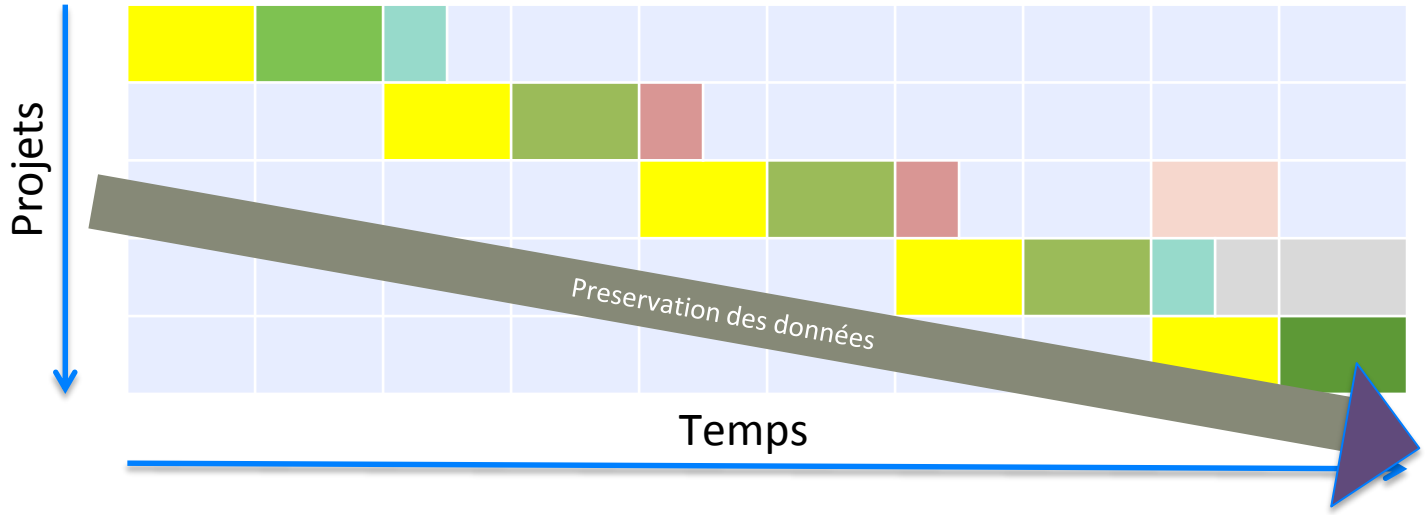


C.Diaconu

Quand faut-il commencer à préserver?

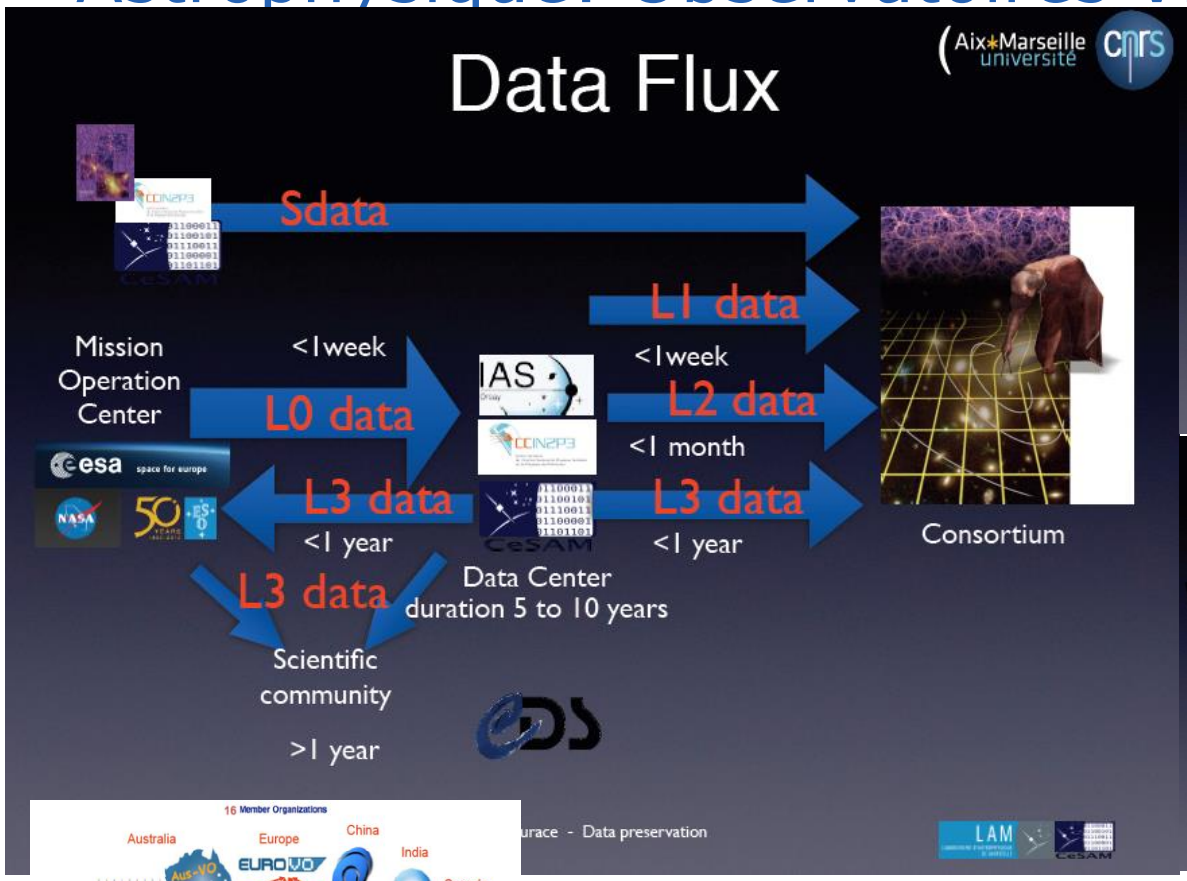


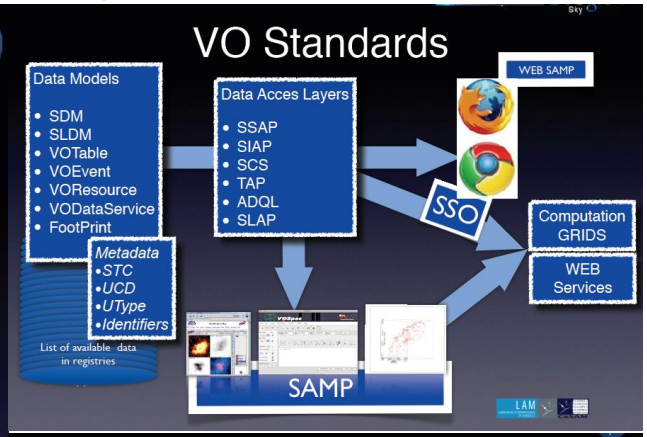
« Data Management Plan » doit inclure la préservation et l'accès à long tèrme



Programme cohérent de la préservation de données

Astrophysique: Observatoires Virtuels





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;

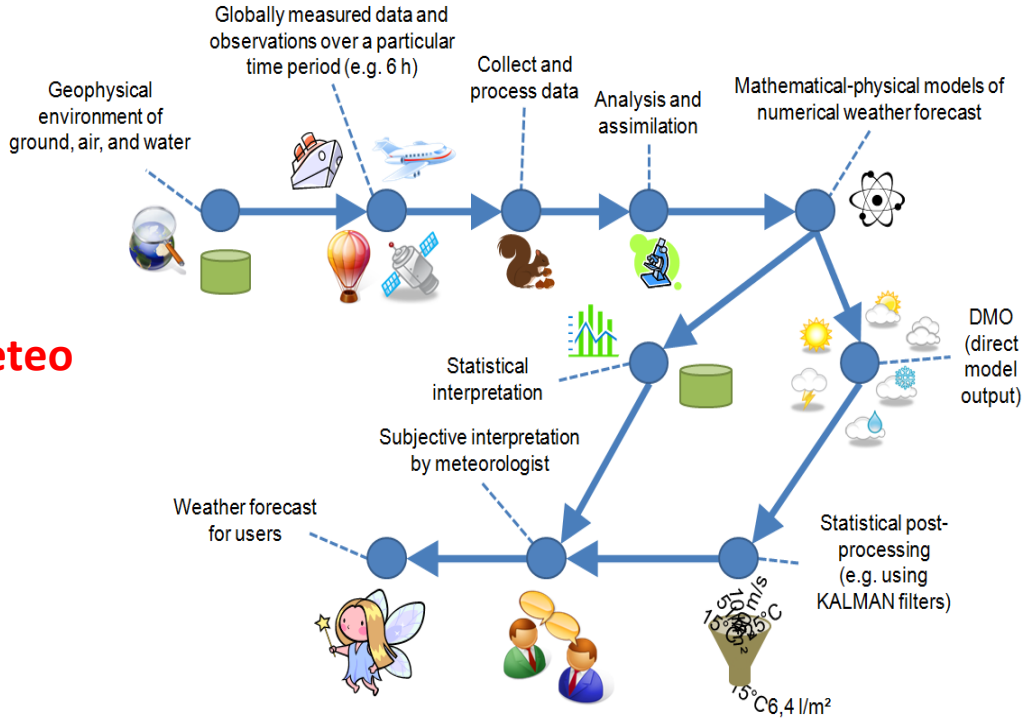
(3) In its report of May 2012 (see Annex 2), the study group provided a concrete proposal for an international collaboration in charge of the Project and data management and policies in high-energy physics;

(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;

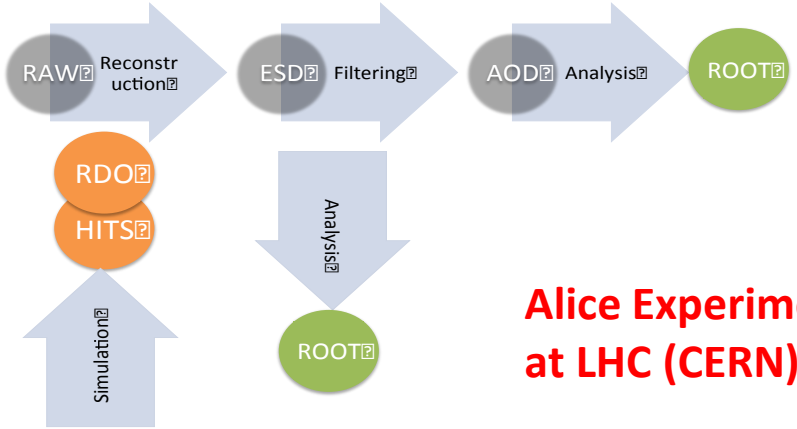
First DPHEP Collaboration Board

Formats, workflows et préservation



Meteo

=



Alice Experiment at LHC (CERN)

Formats de données: standards?

Similarité entre les disciplines

Approche théorique rigoureuse
Besoin et opportunité

Aix*Marseille université
PREDON

http://predon.org

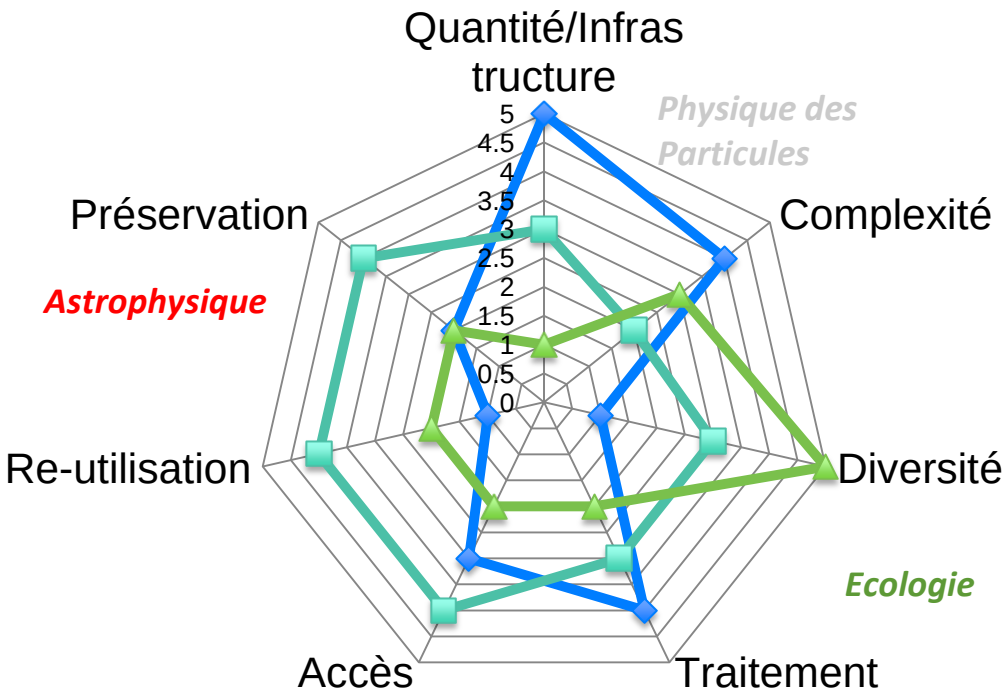
Projet dans le cadre « Mastodons/Big Data » de la MI/CNRS et action dans Madics (GDR Big Data)

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

Harmonisation R&D

Architecture, Pilotage

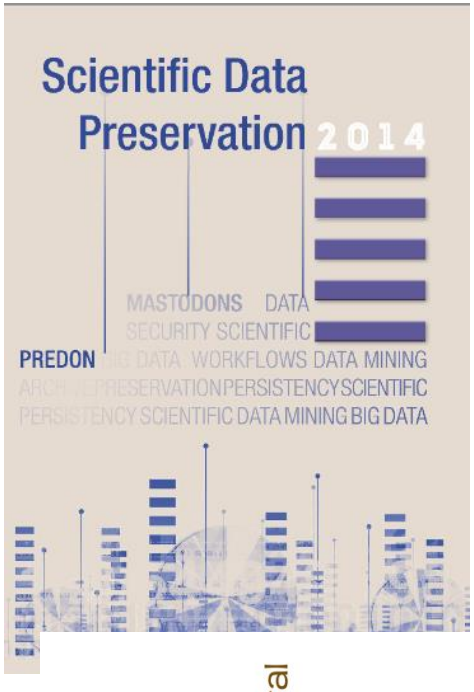
Multidisciplinarité, complémentarité

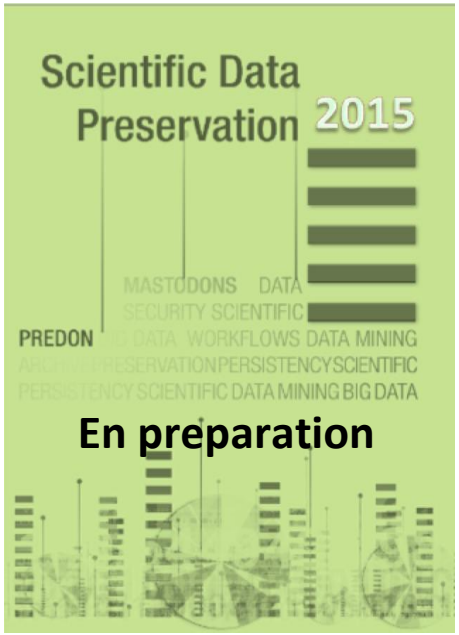


Exemple sur 3 domaines

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

Document PREDON 2014





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

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 reproducibilité en bioinformatique

S. Cohen-Boulakia
Ch. Blanchet
O. Collin

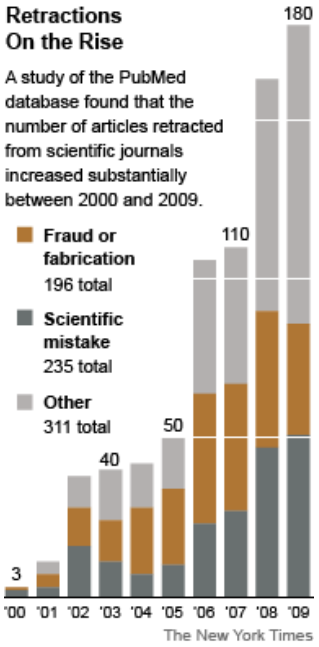
Nekrutenko & Taylor, Nature Genetics (2012)

- **50 articles résultats obtenus avec un outil ➔ 7/50 (14%) reproducibles**

Alsheikh-Ali et al, PLoS one (2011)

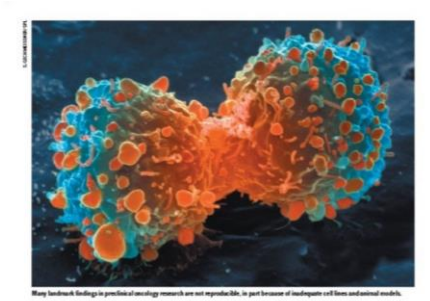
- **10 articles dans le top-50 journaux IF ➔ 47 /500 (9%) reproducibles**

Consequences 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.

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.

Interopérabilité des bases de données en écologie

Romain David

imbe
institut méditerranéen de biodiversité et d'écologie marine et continentale

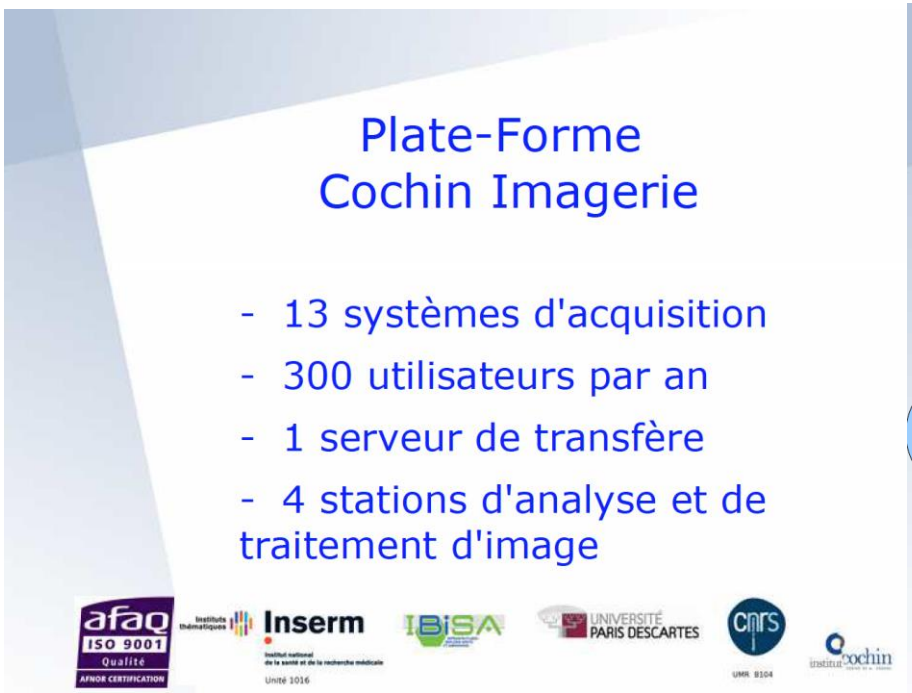
Acquisition automatisée
Systèmes de reconnaissance automatique
Outils d'aide à la détermination
Data management
Indexation de la donnée
Gestion de la qualité et réutilisation de la donnée
Qualification et ontologie
Fouille de données
Conservation des données (Nouveau : les variables essentielles de biodiversité)
Représentations des données
…

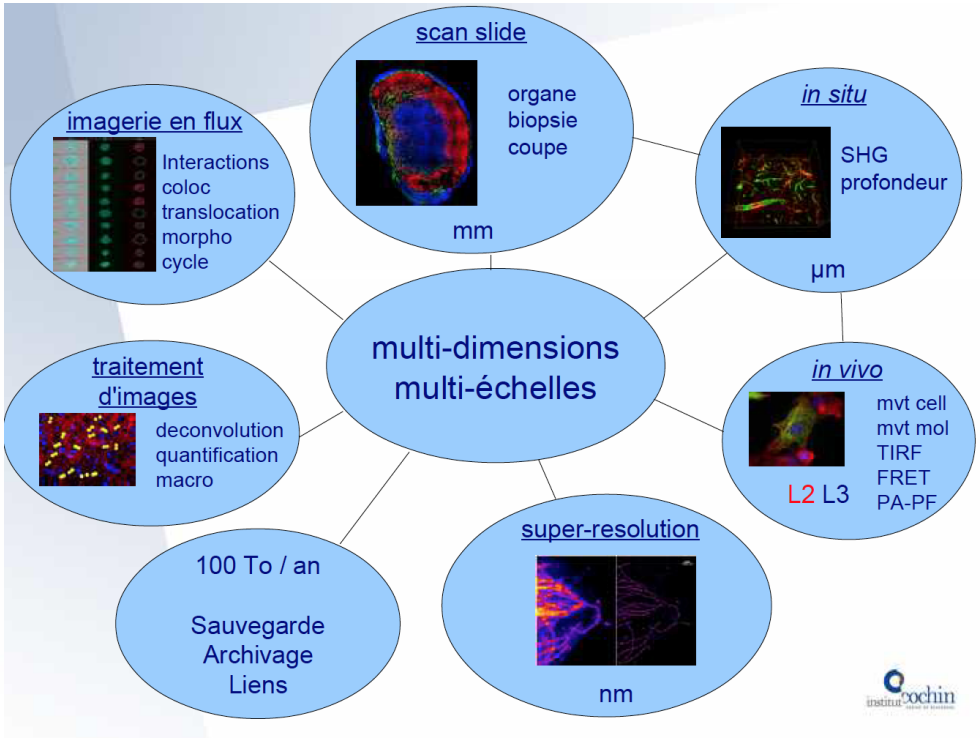
Organisation d'un workshop IndexMED-PREDON en mars 2015

C.Diaconu

Imagerie Cellulaire et le Big Data

P. Bourdoncle





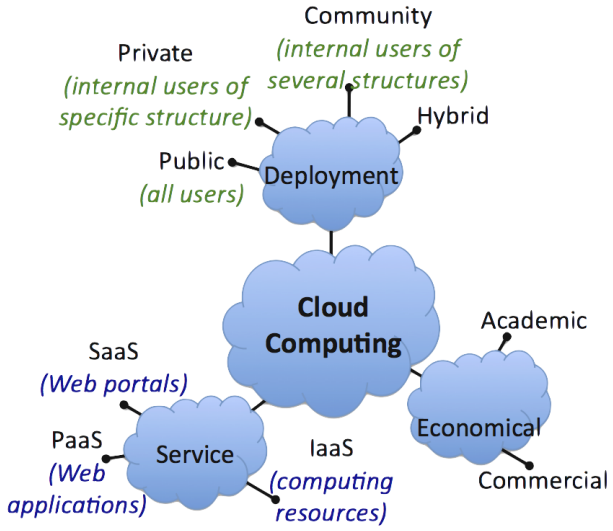
Un potentiel important de coopération:

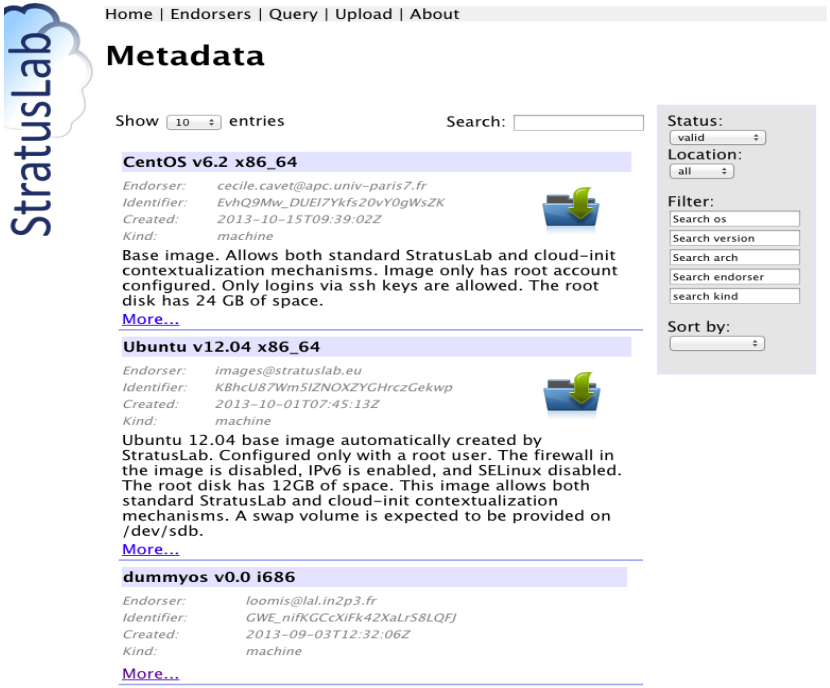
- **Structuration, méthodes, technologies**

C.Diaconu

Préservation de données dans le « cloud »?

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





Exemple: **StratusLab**

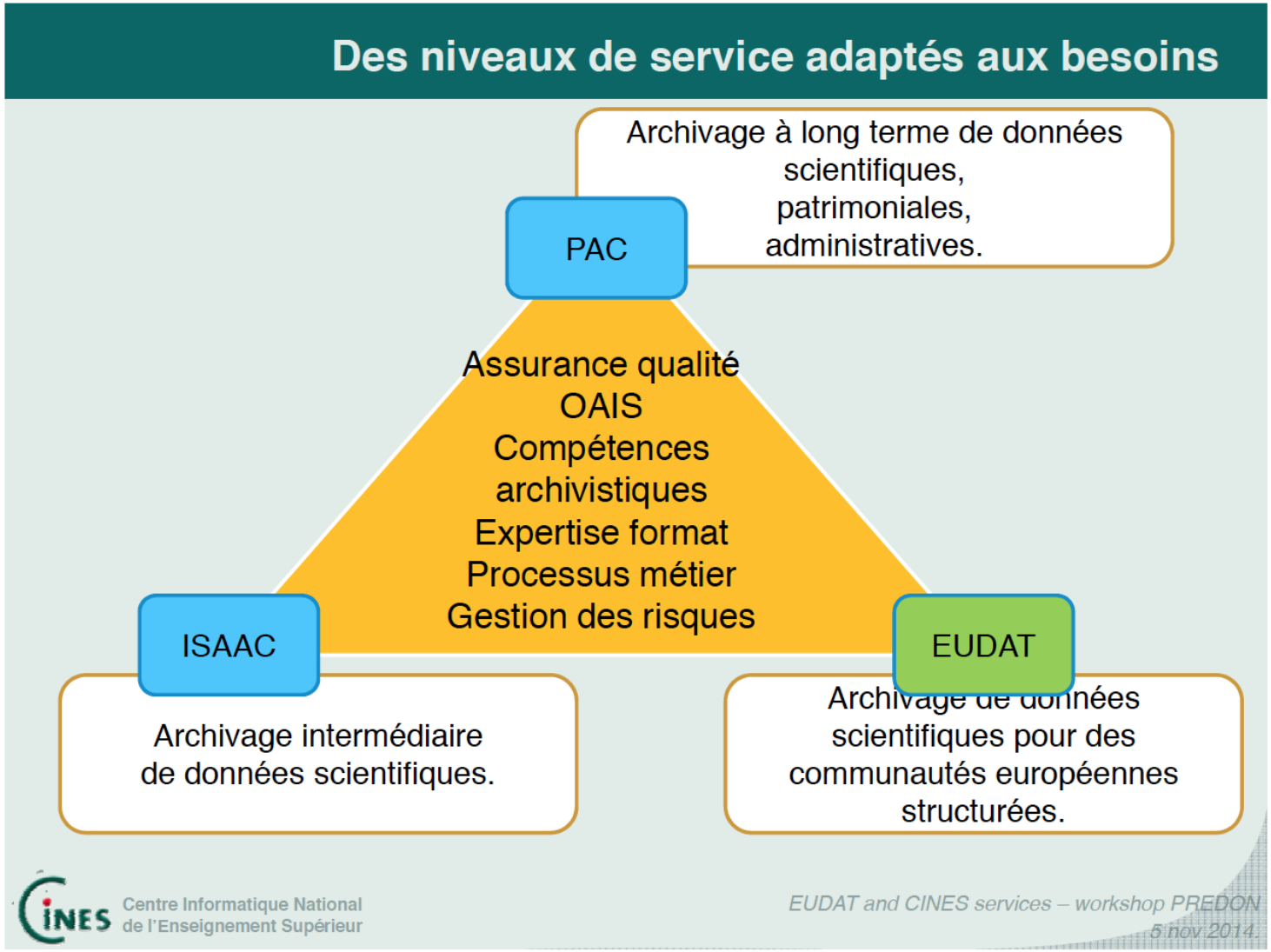
(http://stratuslab.eu/index.html)
End-user client
MarketPlace (OS collection)
Persistent disk Web interface
Ressource monitoring

Disk images have **6 months of validity**
OS update/upgrade for security.

Virtualisation/cloud need to be tuned for long term

C.Diaconu

Projets CINES



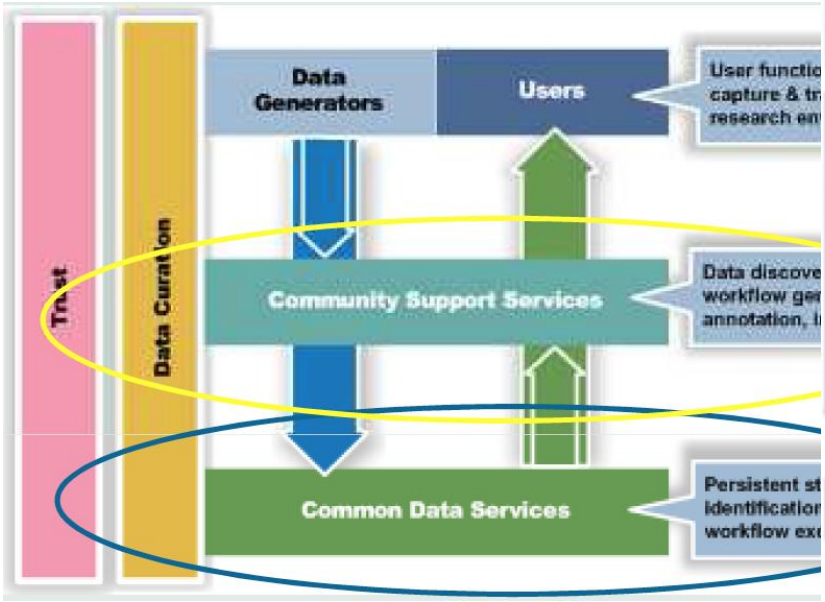
C.Diaconu

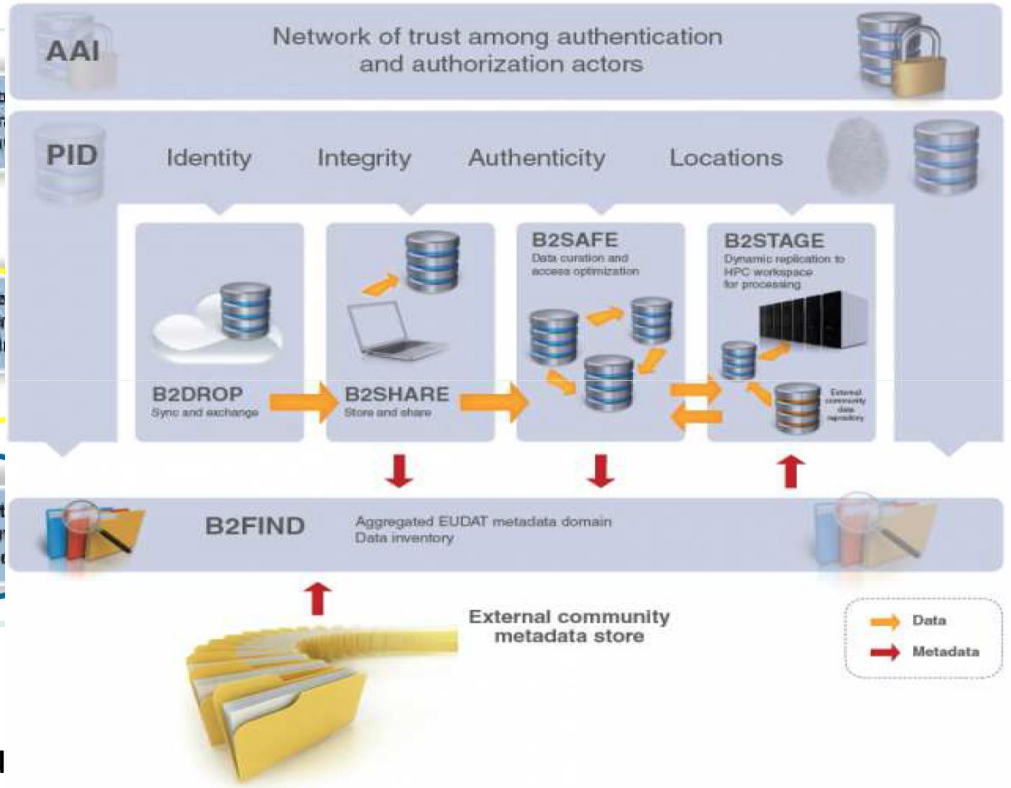
EUDAT

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 serviceinfrastructure when they are of use to other communities.**



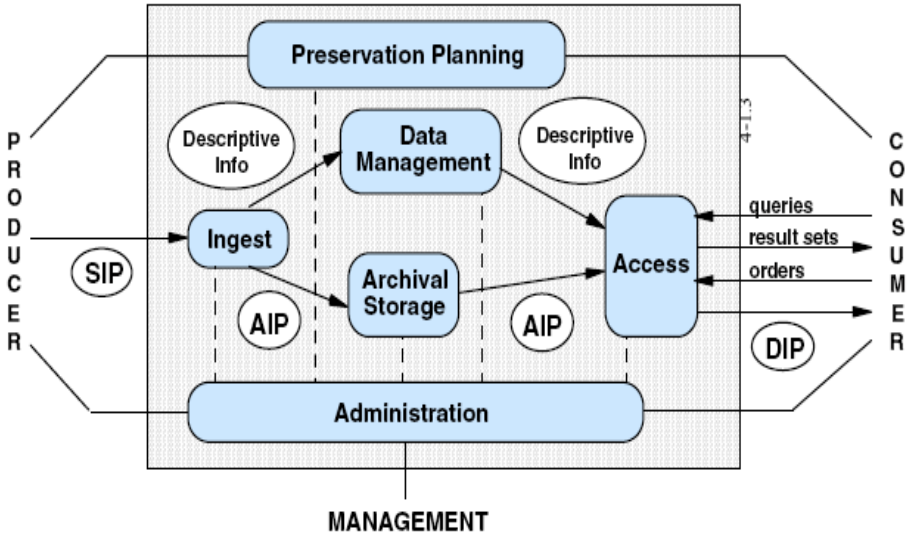


C.Diaconu

Open Archive Information System OAIS

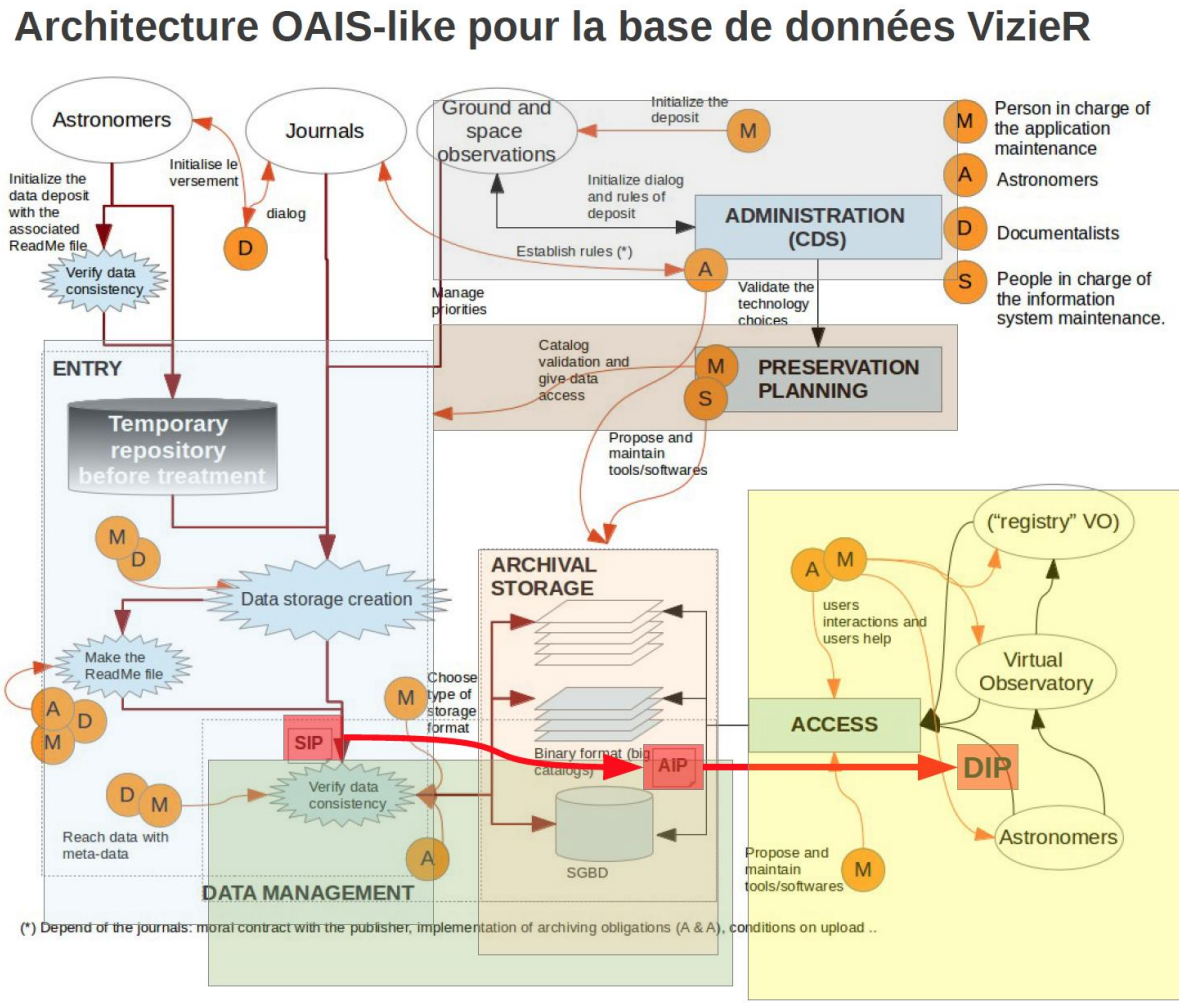
OAIS = Modèle conceptuel et fonctionnel destiné à la gestion, l'archivage et à la préservation à long terme de documents numériques.

- **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



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

Search

- 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