La préservation des données scientifiques: une mine d'or pour la science de demain

Workshop PREDON

APC 5-6 Novembre 2014

C. Diaconu

Centre de Physique des Particules de Marseille

CNRS et Aix Marseille Université (AMU)

Les données digitales sont fragiles

- La capacité de stockage est physiquement dépassée depuis longtemps

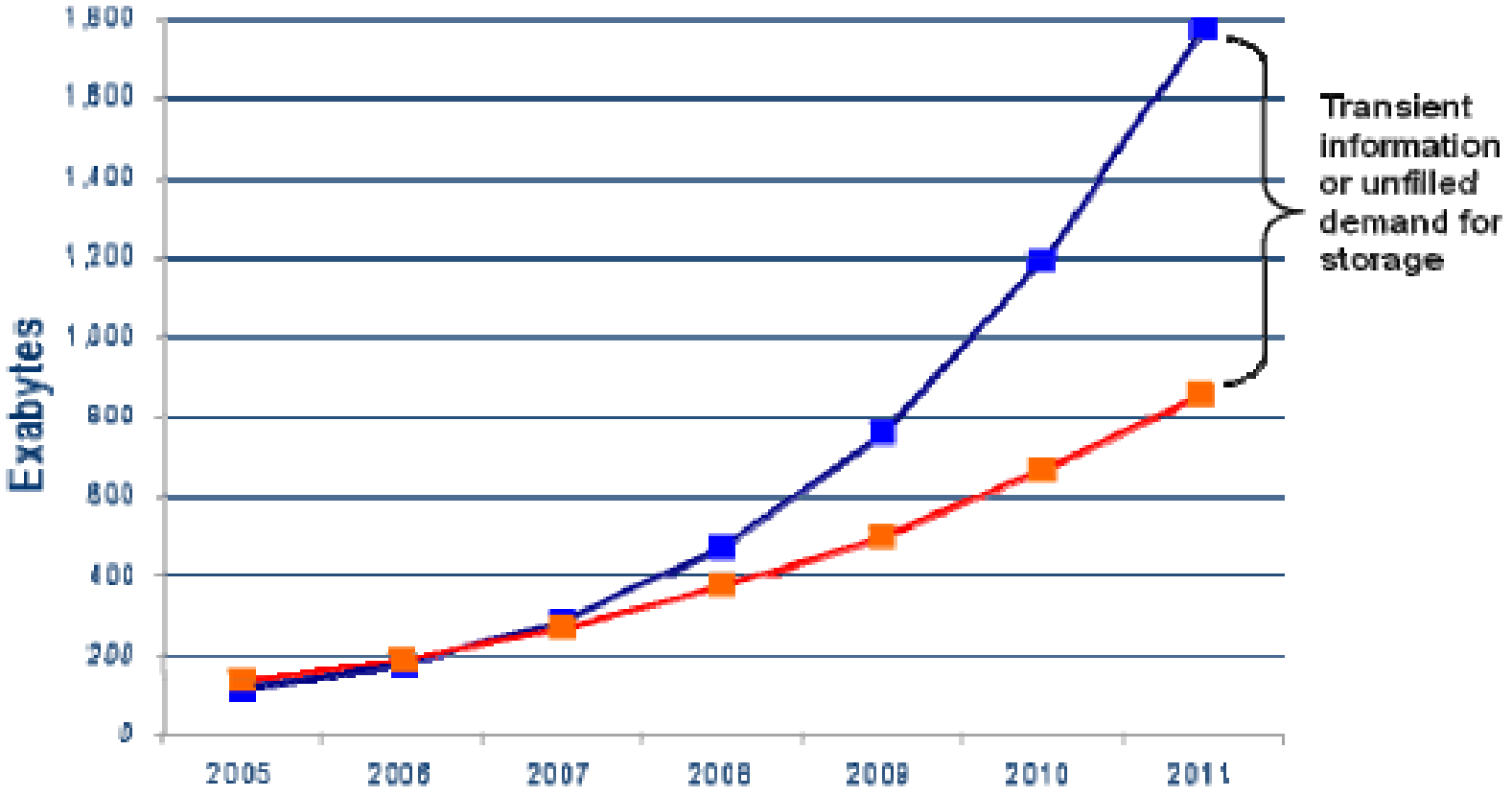
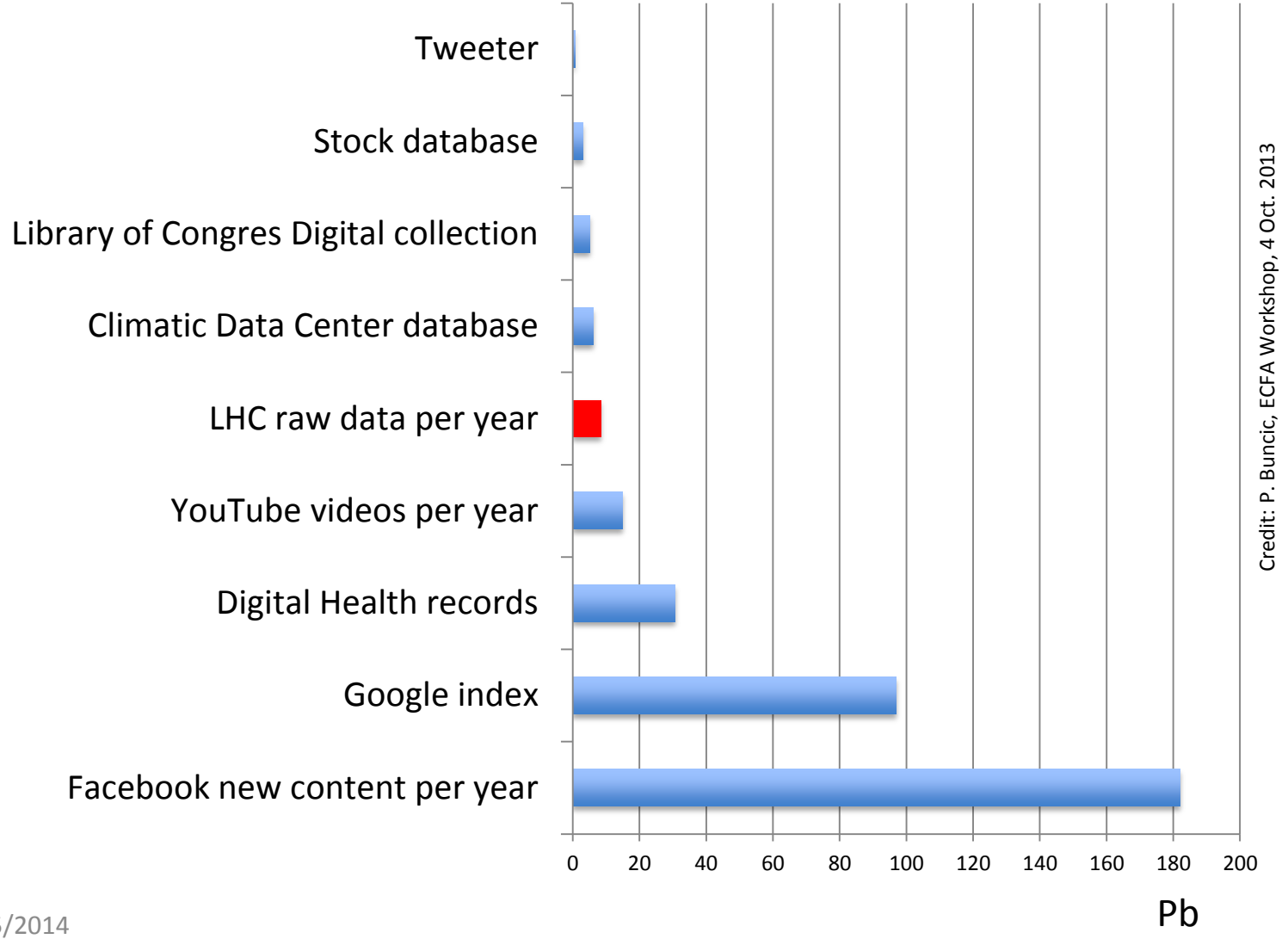


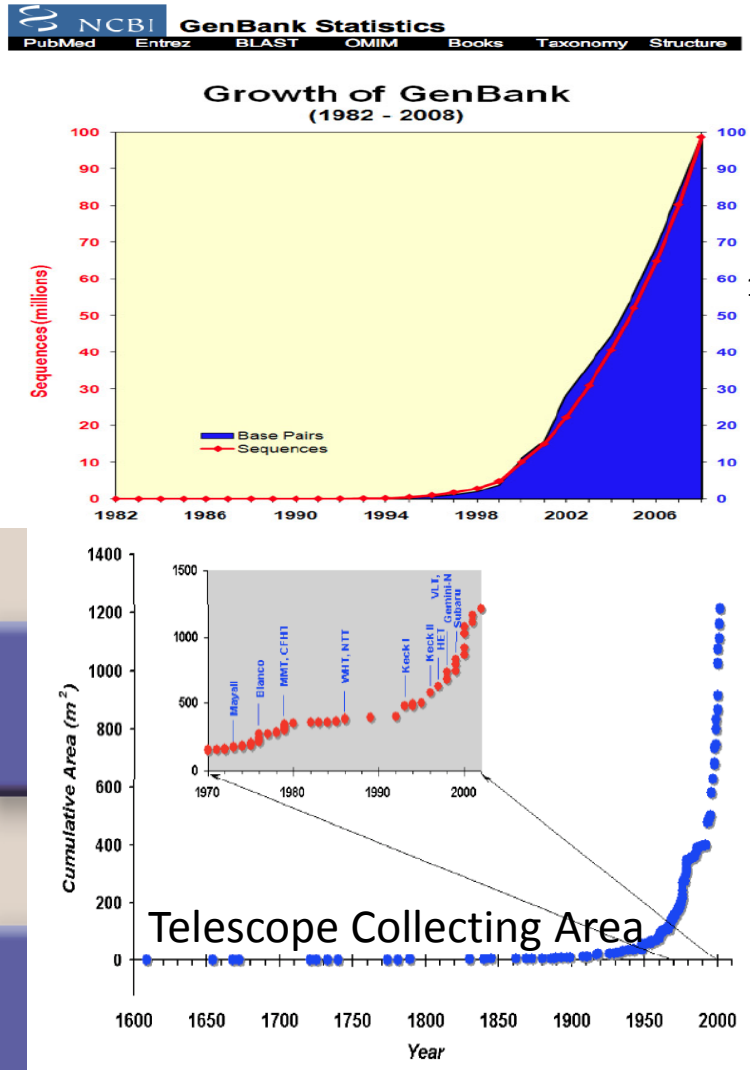
FIGURE 1.3: **Information and Storage**
Source: J. Gantz January 2008 (revised). Used with permission.

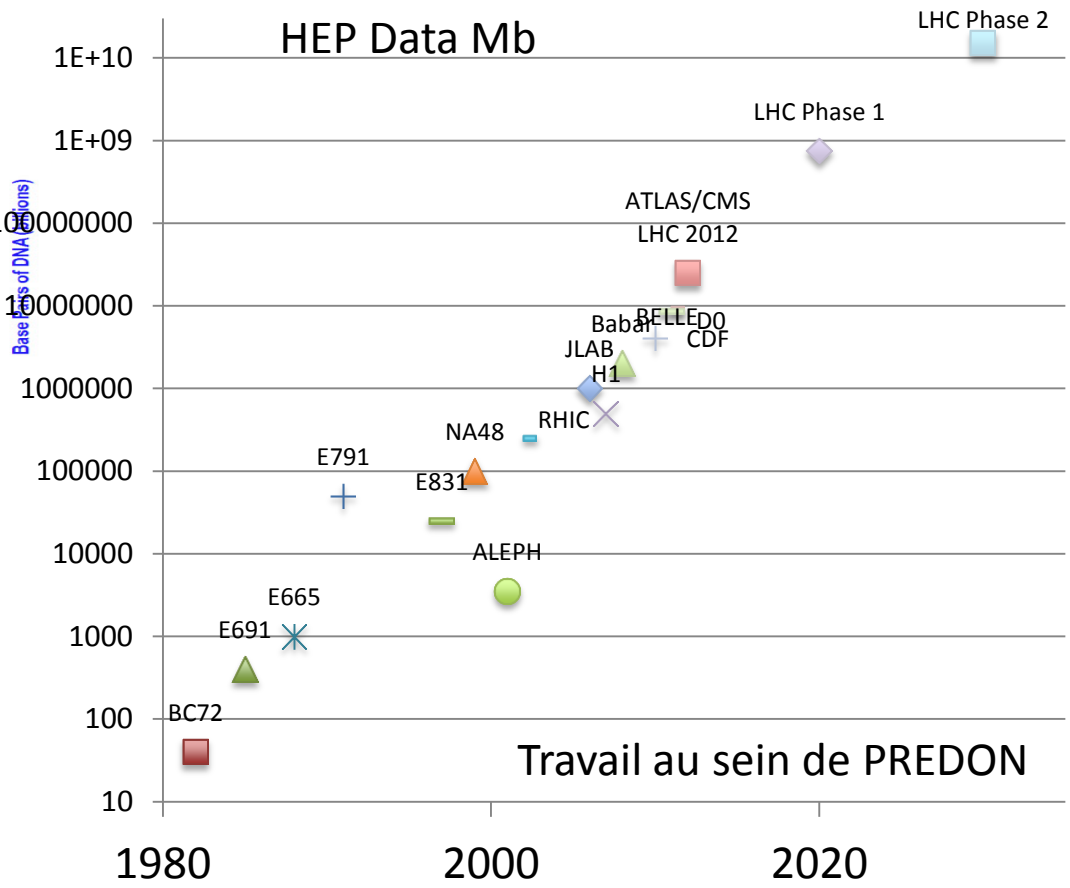
Données digitales explosent (les données scientifiques aussi)



« Big Scientific Data »

- La recherche est « digitale »
  - Augmentation dramatique de la quantité/complexité des données

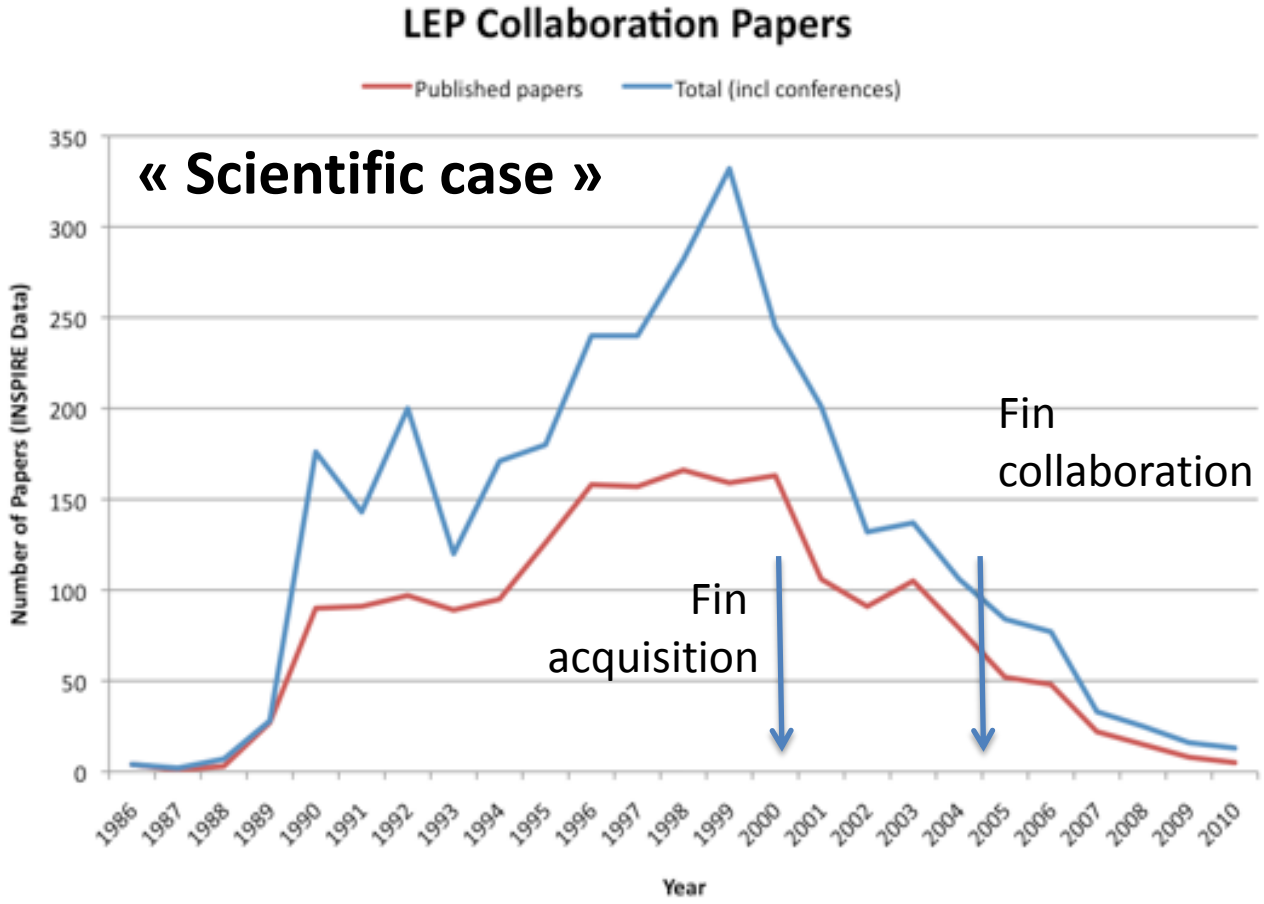


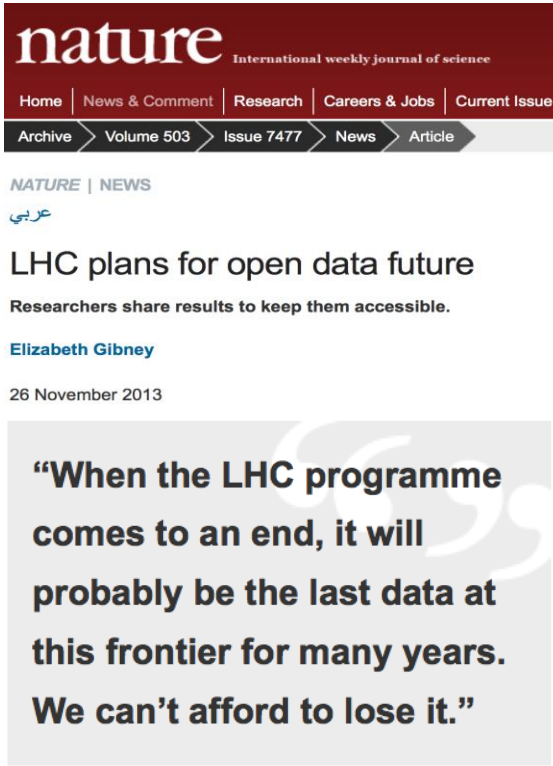


Est-ce que les données scientifiques sont spéciales (« big » à part)?

- Riches en information
  - structurées suivant un plan de recherche et une démarche scientifique
- De plus en plus diverses
  - la plupart des disciplines produisent massivement des données
- Souvent produites avec des efforts financiers et humains significatifs (voir gigantesques)
  - Plus ça coute cher, moins c'est reproductible
- Englobent des connaissances uniques
  - « Time stamped »
- De plus en plus dans une logique « observatoire »:
  - Les données contiennent plus que ce qu'on voulait au départ
  - Seulement l'information décantée est publiée de suite (1/10)
- PRESERVATION!

Est-ce que ça vaut le coup de garder des données « anciennes »?



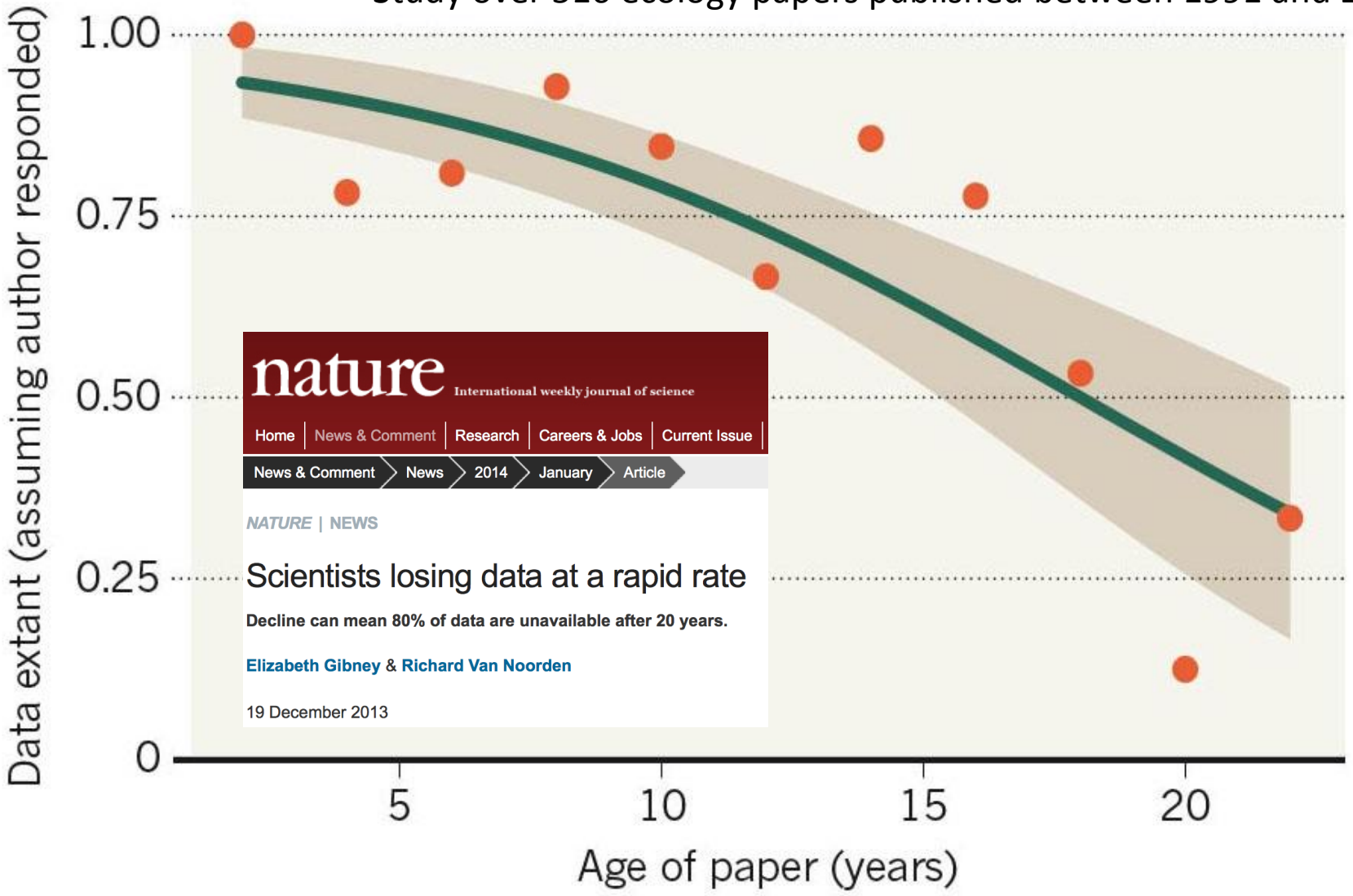


Estimation: gain scientifique de 10% pour un cout bien inférieur à 1%

MISSING DATA

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

Study over 516 ecology papers published between 1991 and 2011.



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue

News & Comment > News > 2014 > January > Article

NATURE | NEWS

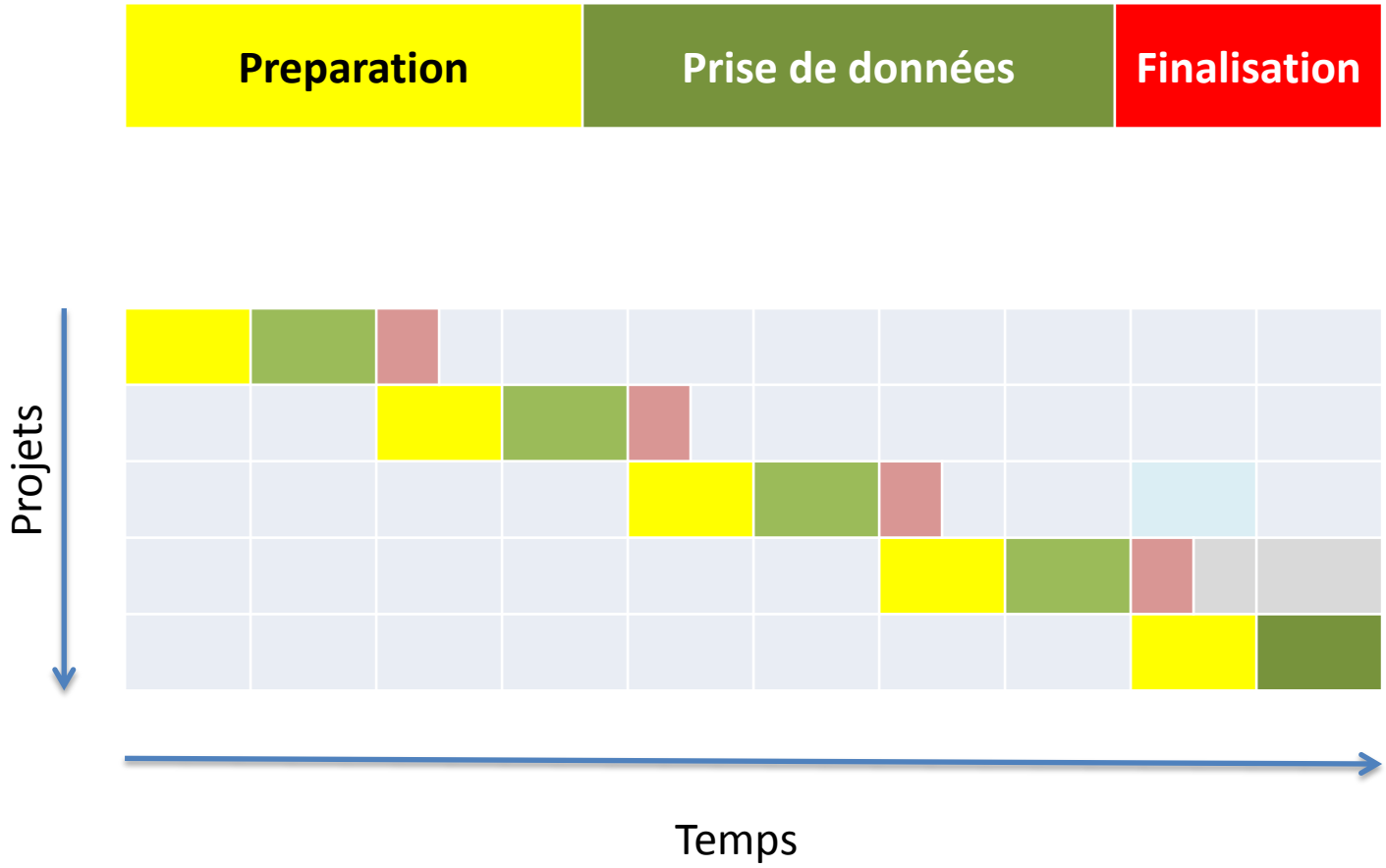
Scientists losing data at a rapid rate

Decline can mean 80% of data are unavailable after 20 years.

Elizabeth Gibney & Richard Van Noorden

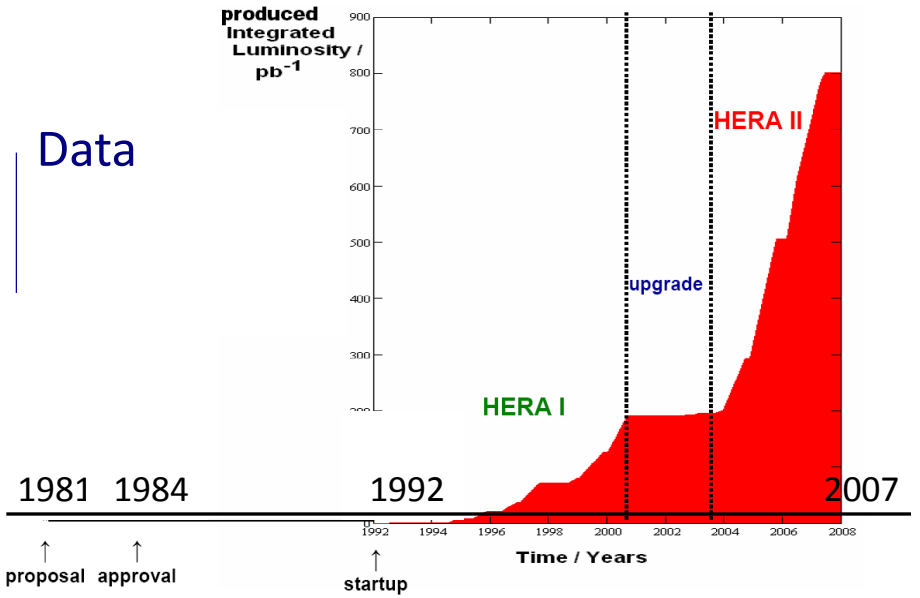
19 December 2013

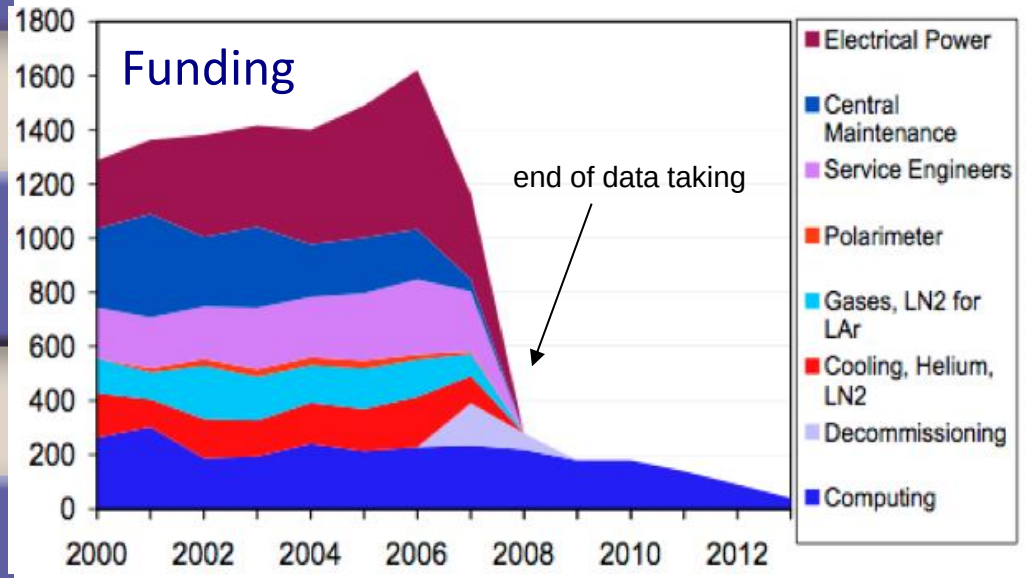
Pourqoui ca pose un problème?

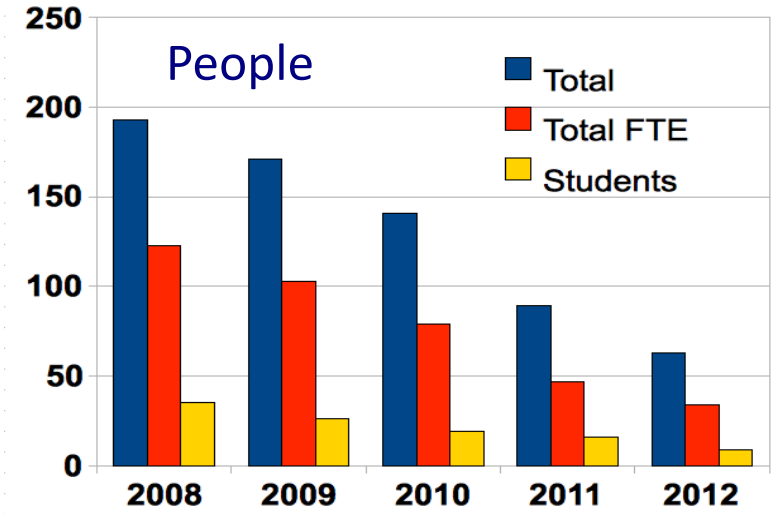


Le contexte de la préservation de données

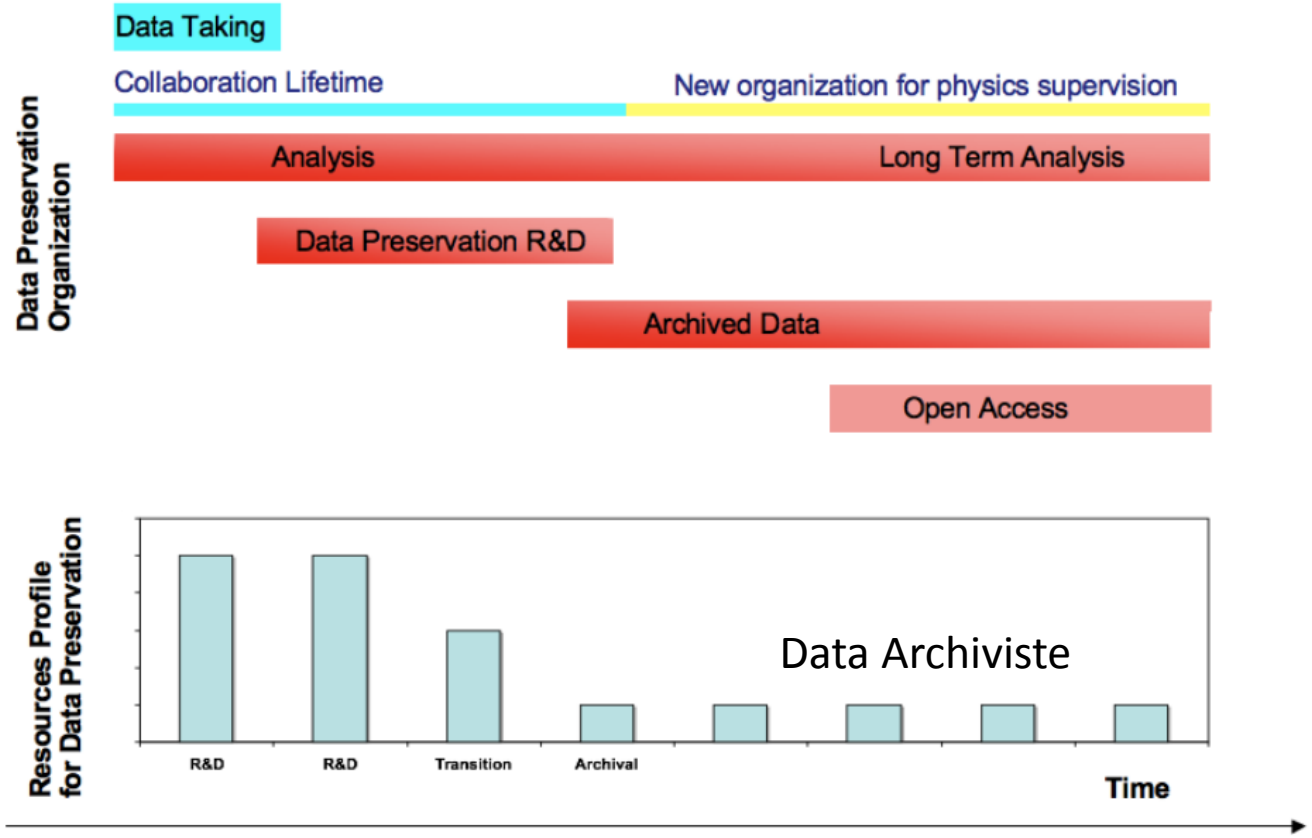
- La fin des programmes scientifiques est le plus gros danger pour les données
  - Décrue des resources et du personnel
  - Organisation a long terme est cruciale





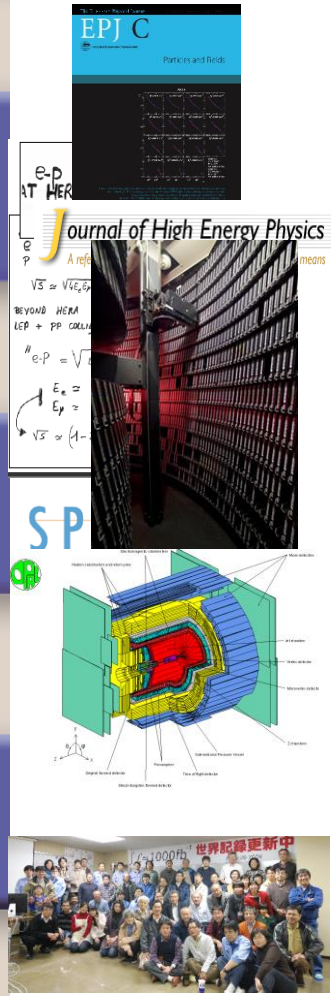


Préservation: modèle économique te organisationnel



The specific costs around 1% of the project
Scientific outcome around 10% more papers

Données Scientifiques



- Publications
- Documentation
- Raw
- Donées Processées
- Meta-données
- Workflows
- Software
- Diffuse knowledge

….more…

Complexité, couts
Utilité

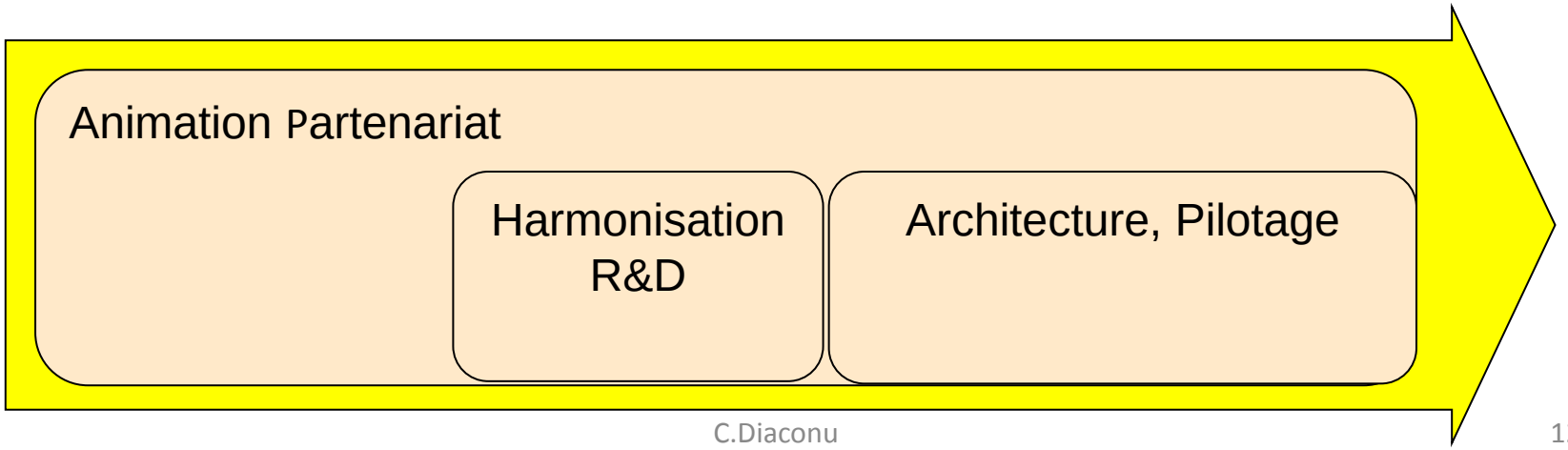
Technologie, méthodologie Organisation

Quel modèle de préservation pour les données scientifiques?

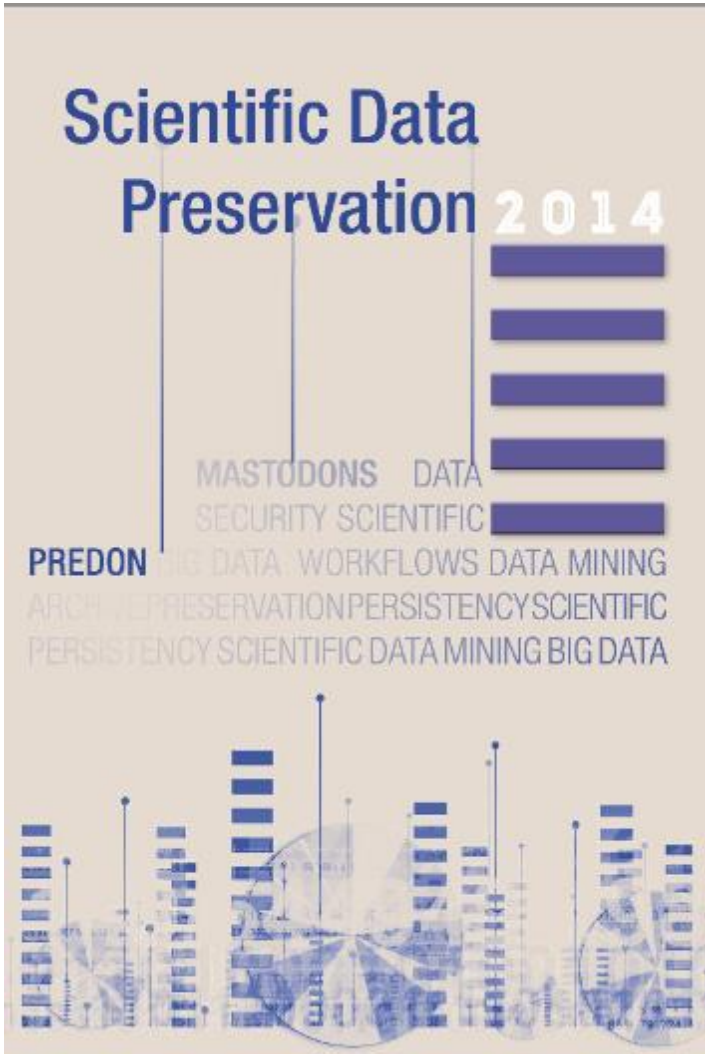
PREDON http://predon.org

- Projet dans le cadre « Mastodons/Big Data » de la MI/CNRS

	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	+	++	+++	+	+++



Livre blanc sur la préservation de données (« facts finding »)

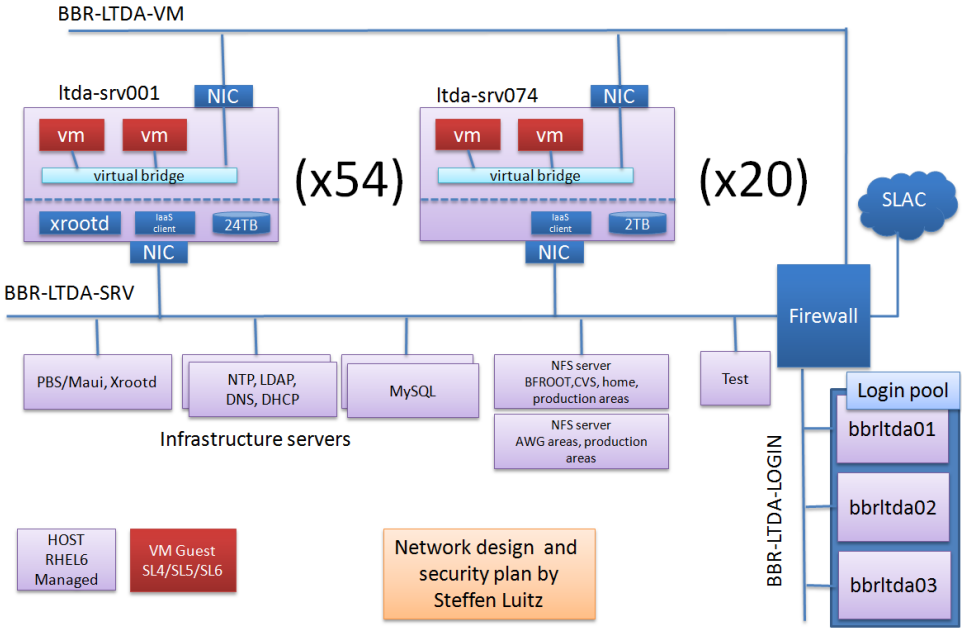


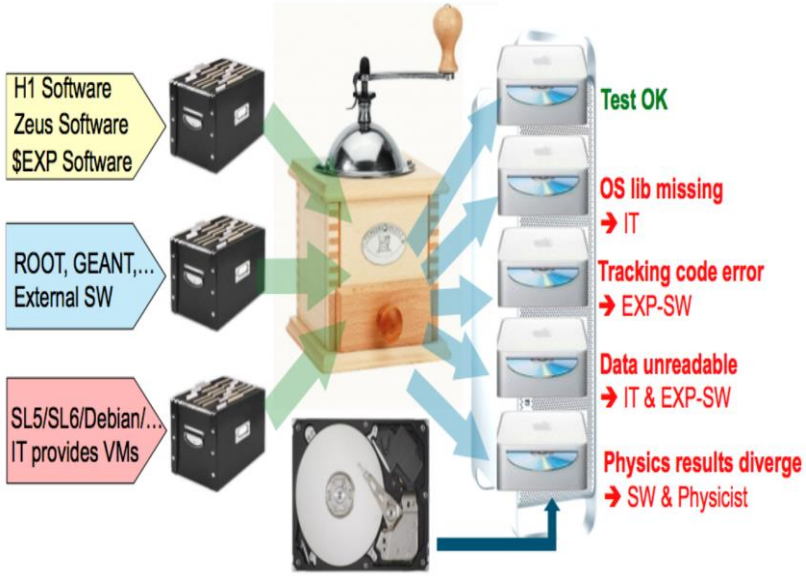
Physique des Particules

dphep.org

Préservation d'un système d'accès et calcul à des données complexes (SLAC/Stanford USA)

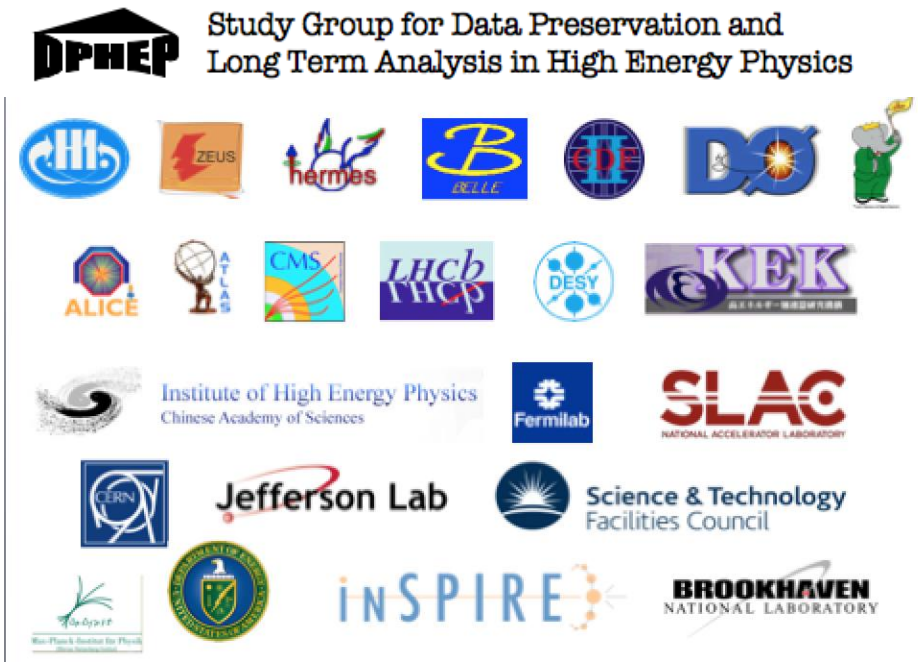
Système de préservation et migration Virtualisation, validation intensive (DESY, Hambourg, Allemagne)

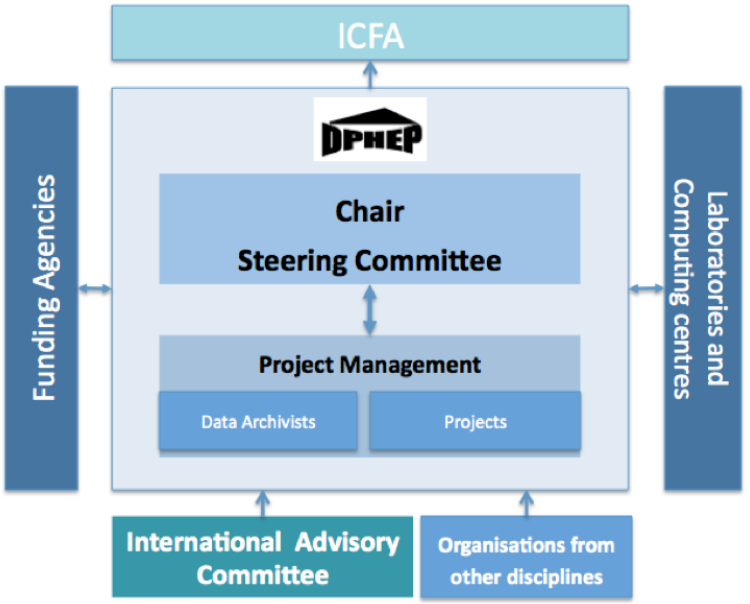




DPHEP: « Project Manager » nommé au CERN en Octobre 2012 (Scientific chair: CD)

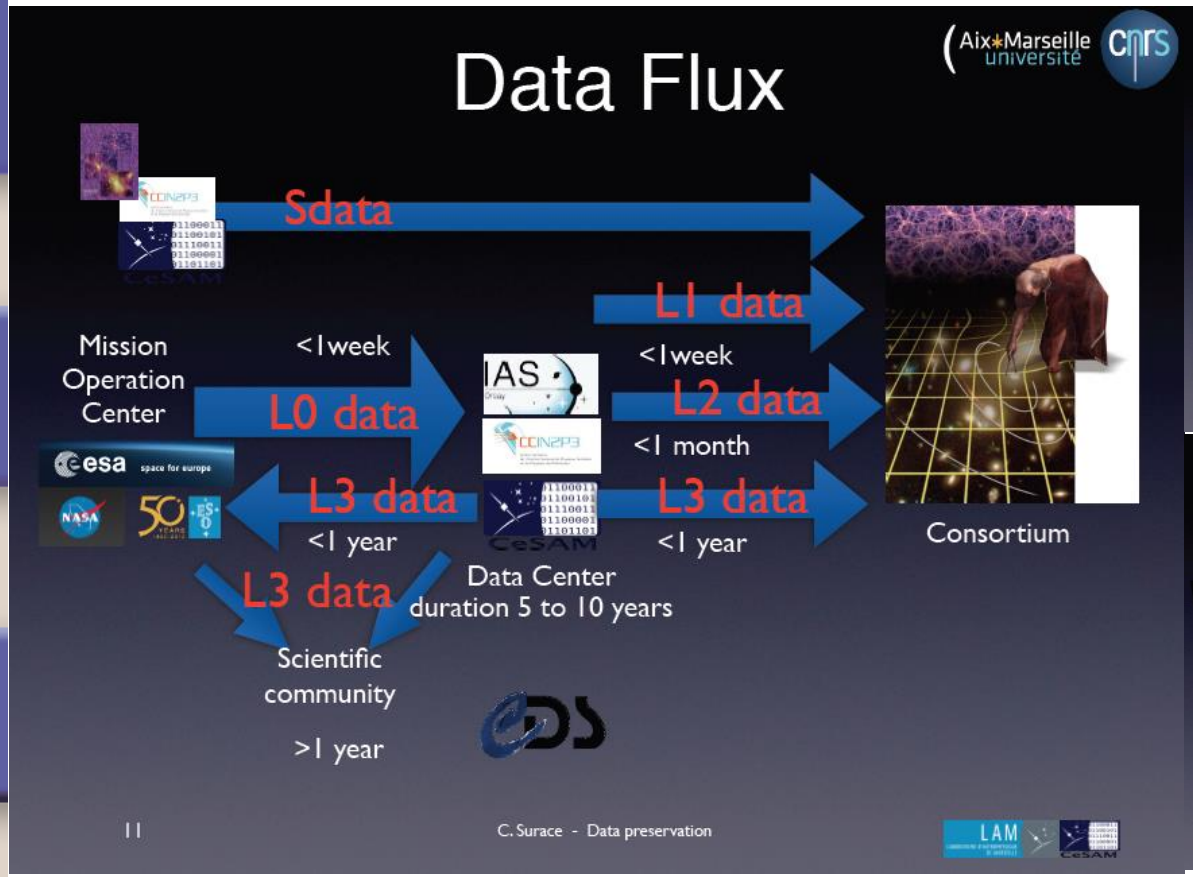
DPHEP:International organisation

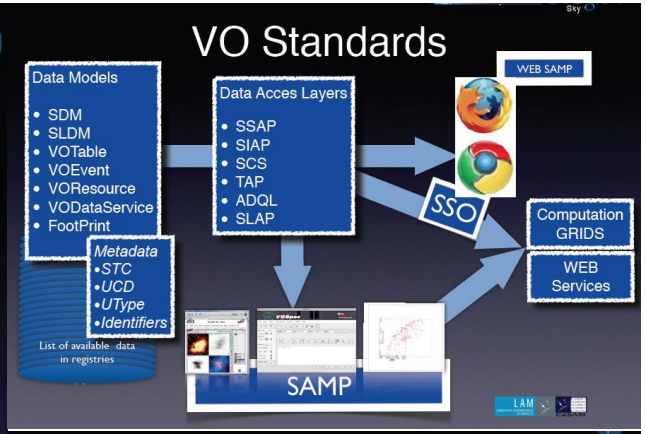




- Study Group DPHEP:
  - Large laboratories CERN, DESY, FERMILAB, SLAC, KEK,IHEP and experiments
- **Breaking news: Organisation internationale mise en place**
  - MoU signé en Juillet 2014
  - 100 contact personnes de contact
  - Chair: D. Diaconu Project Manager: Jamie Shiers (CERN)

Exemple projet astrophysique: Virtual Observatories

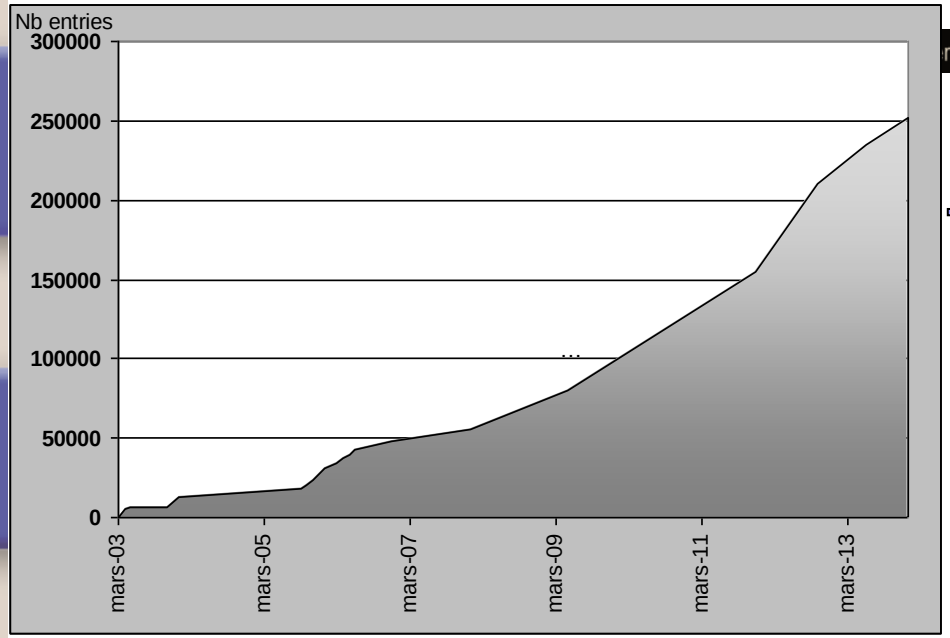


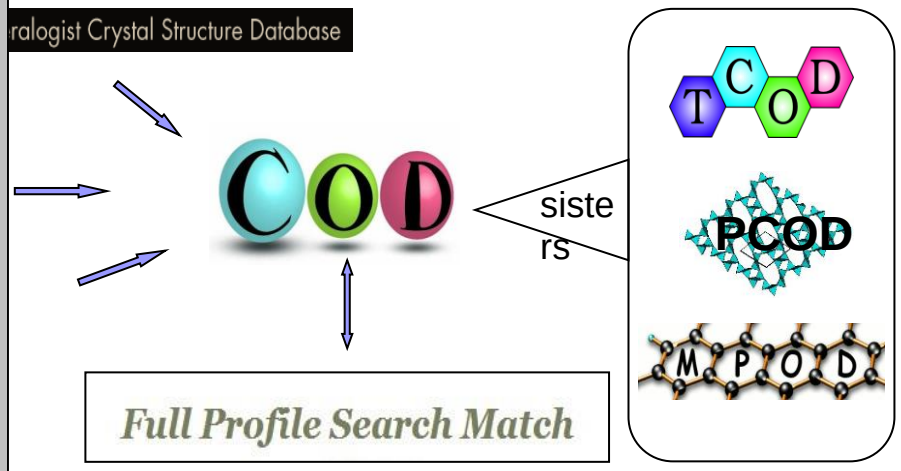


Nouveau Contact 2013

Crystallography Open Databases and Preservation: a World-Wide Initiative

Daniel Chateigner (for the COD Advisory Board)





"…there **is not yet sufficient coherence** of experimental metadata standards or national policy to rely on instrumental facilities to act as permanent archives;

-there **is not sufficient funding** for existing crystallographic database organisations (which maintain curated archives of processed experimental data and derived structural data sets) to act as centralised stores of raw data, although they could effectively act as centralised metadata catalogues;

-few institutional data repositories yet have the expertise or resources to store the large quantities of data involved with the appropriate level of discoverability and linking to derived publications."

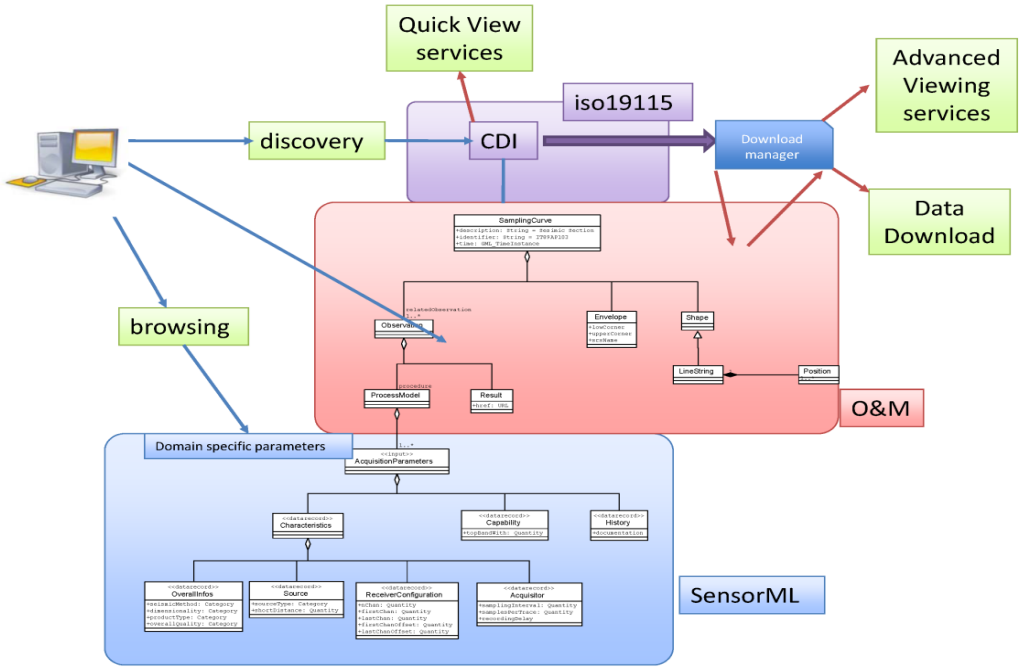
Nouveau Contact 2013

Seismic Data Preservation

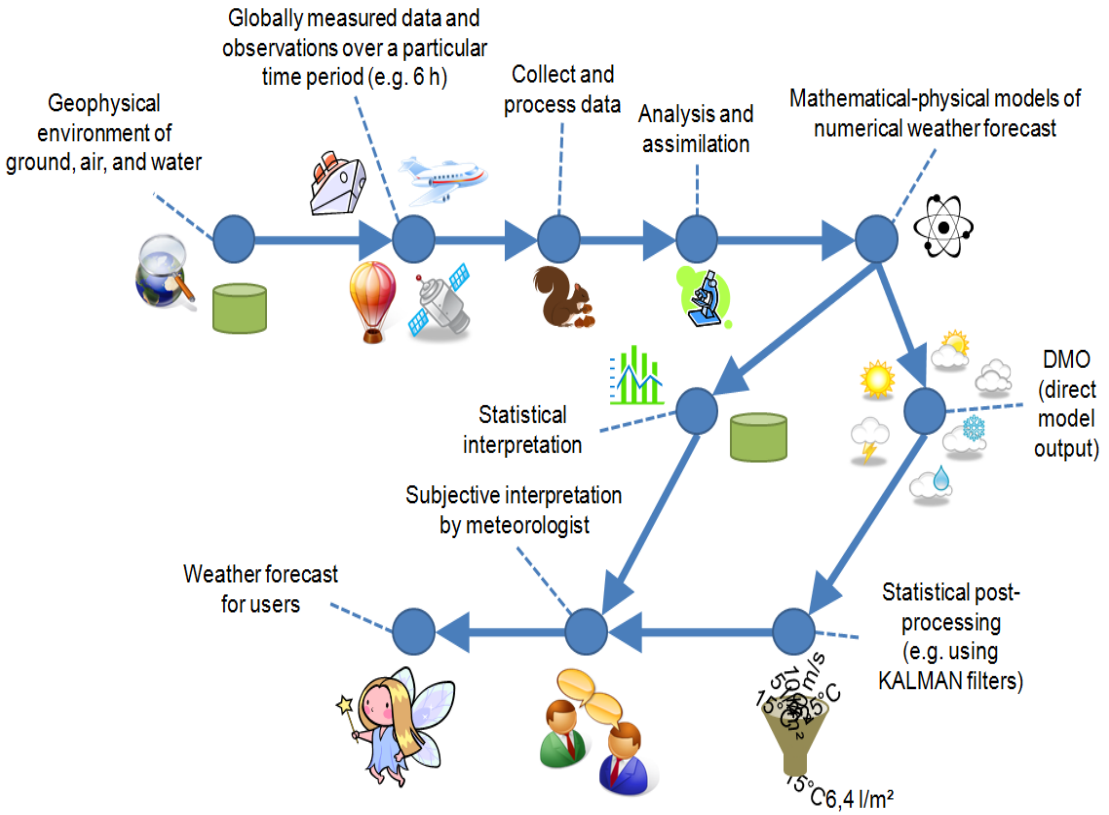
Marc SCHAMING, Institut de Physique du Globe (CNRS/UNISTRA), Strasbourg

Conclusion

Preservation of seismic data is essential, but usually not considered by scientists, because it takes resources to document metadata, to read and copy tapes, to convert formats, etc. These tasks should be addressed at national and/or European level. Some European projects (Seiscan/Seiscanex, Geo-Seas) demonstrated that it is possible and useful. Repositories at national level should pursue this task with geophysical skills.

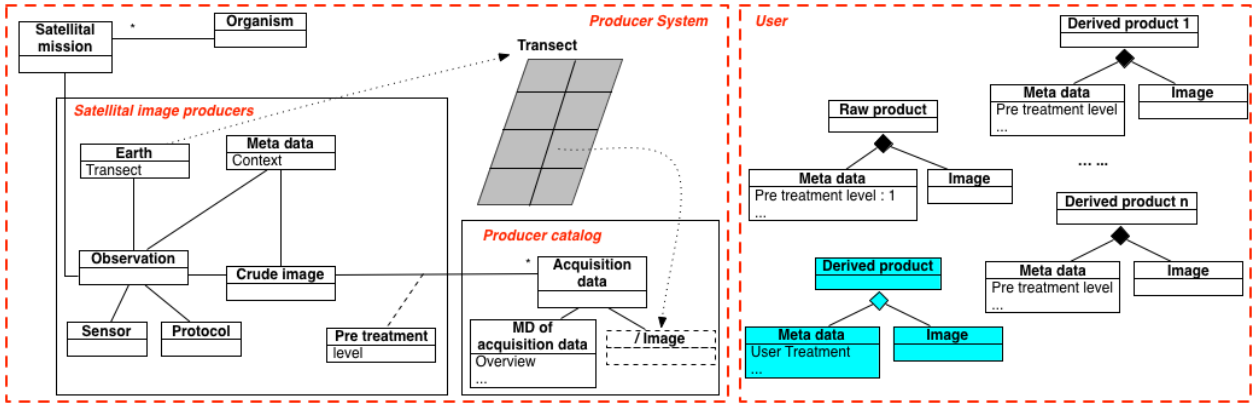


Workflows et préservation



Similarité entre les disciplines

Besoin d'une approche théorique rigoureuse



Nouveau Contact 2013

Long Term Archiving and CCSDS standards

Danièle Boucon, CNES

The primary objective of the Producer-Archive Interface Specification (PAIS) standard is to provide concrete XML files supporting the description and the control of transfers from a Producer to an Archive.

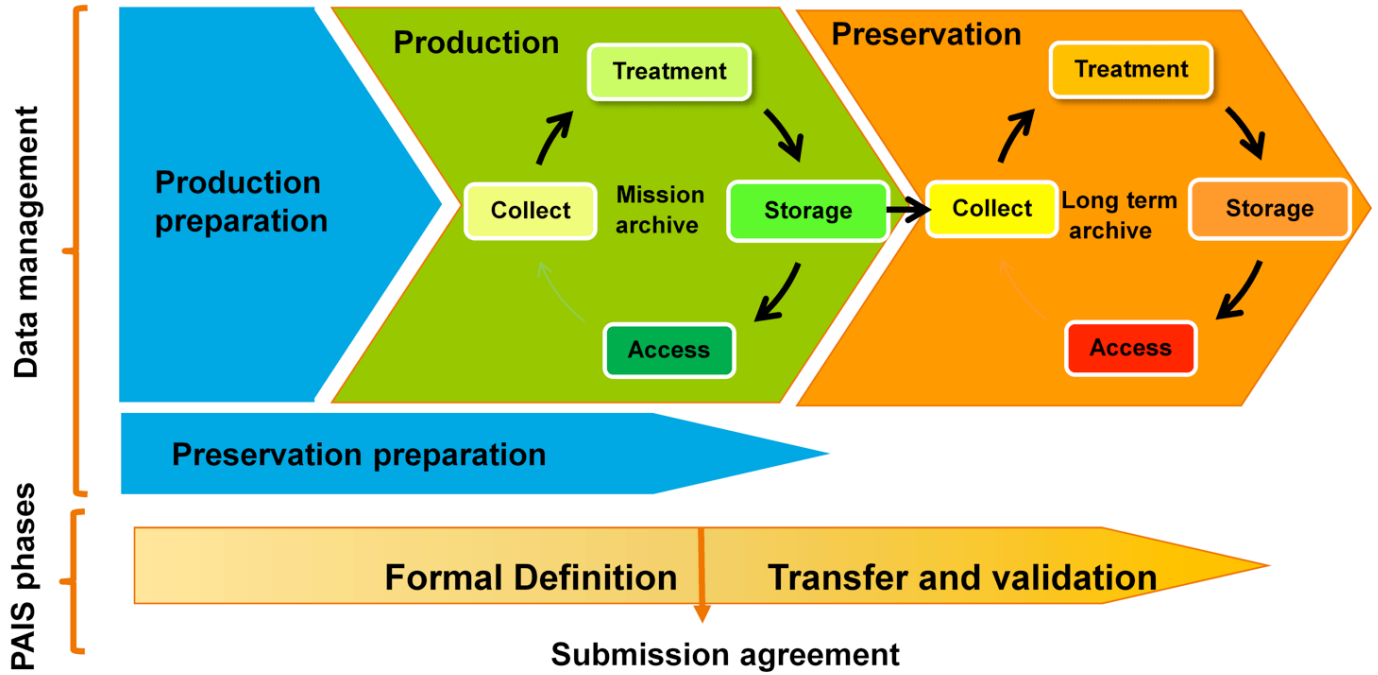


Figure 3: PAIS, preservation process and data lifecycle

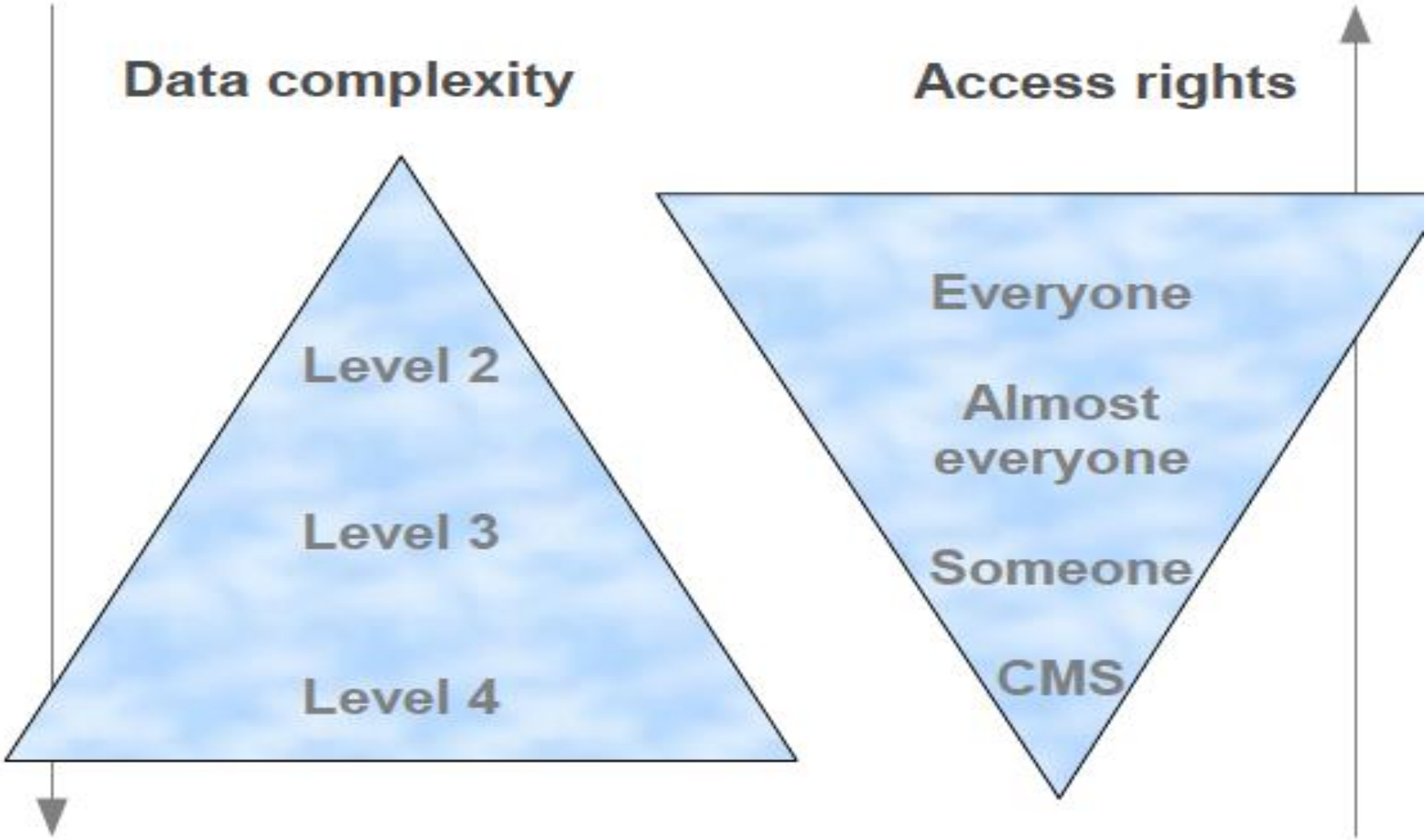
Nouveau Contact 2013

Scientific Data Preservation, Copyright and *Open Science*

Philippe Mouron, Aix-Marseille University, Faculté de droit et de science politique

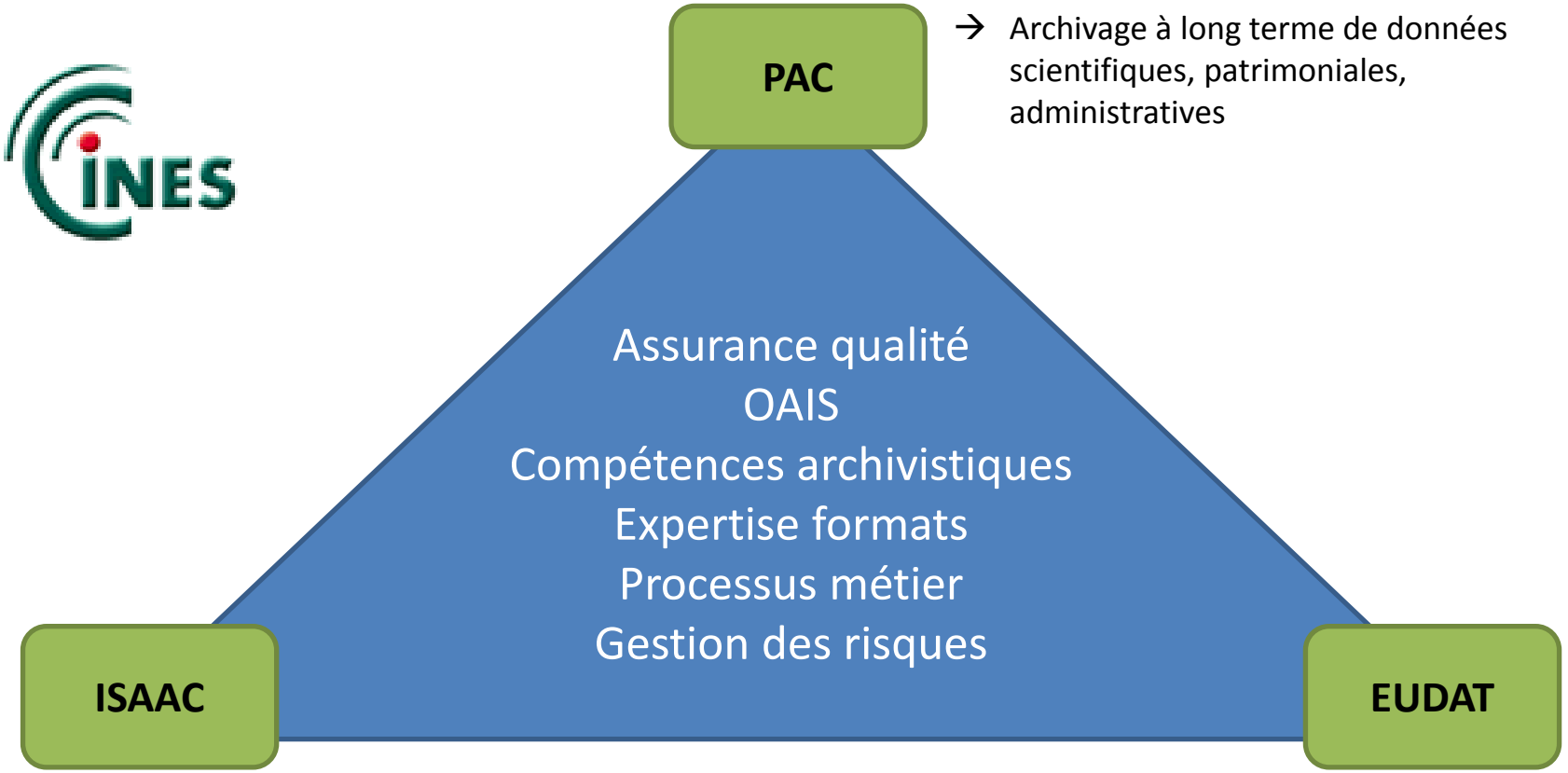
- **The best guarantee for ensuring the integrity of a resource is based on property.**
- However, isn't there a public ownership of scientific research?
  - In truth, even if the public authorities may fundamentally participate in the scientific research, this does not mean, *ipso facto*, that they own its results.
- …any paper, article, report, record, thesis, book, graphic, map,... conducting personal choices of a researcher, or expressing his own personality, will be considered as a work of mind […] are copyrightable
- **The goal of digital preservation of scientific data must therefore be reconciled with intellectual property rights.**
- Open model of management of intellectual property rights.
  - Tools: open access licensis (e.g. Creative Commons)

Preservation complexity levels and access



Archival expertise CINES

Les services d'archivage au CINES



→ Archivage à long terme de données scientifiques, patrimoniales, administratives

→ Archivage intermédiaire de données scientifiques

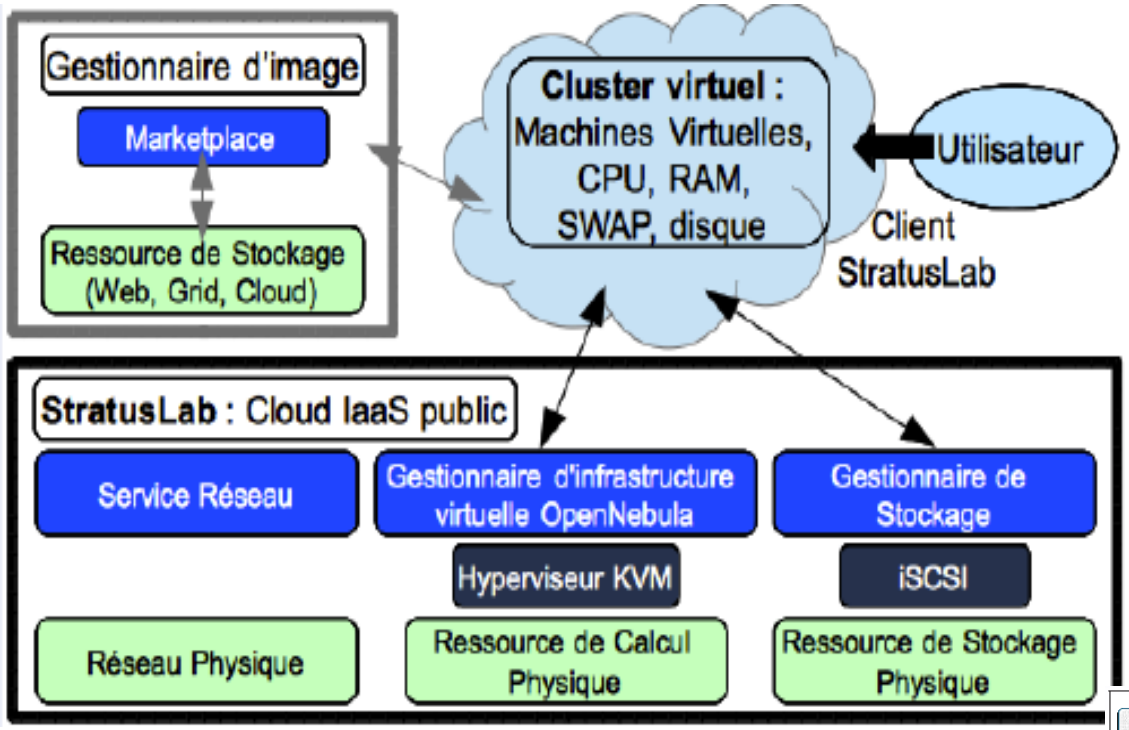
→ Archivage de données scientifiques pour des communautés européennes structurées

Exemple projet: Data processing & storage in the cloud

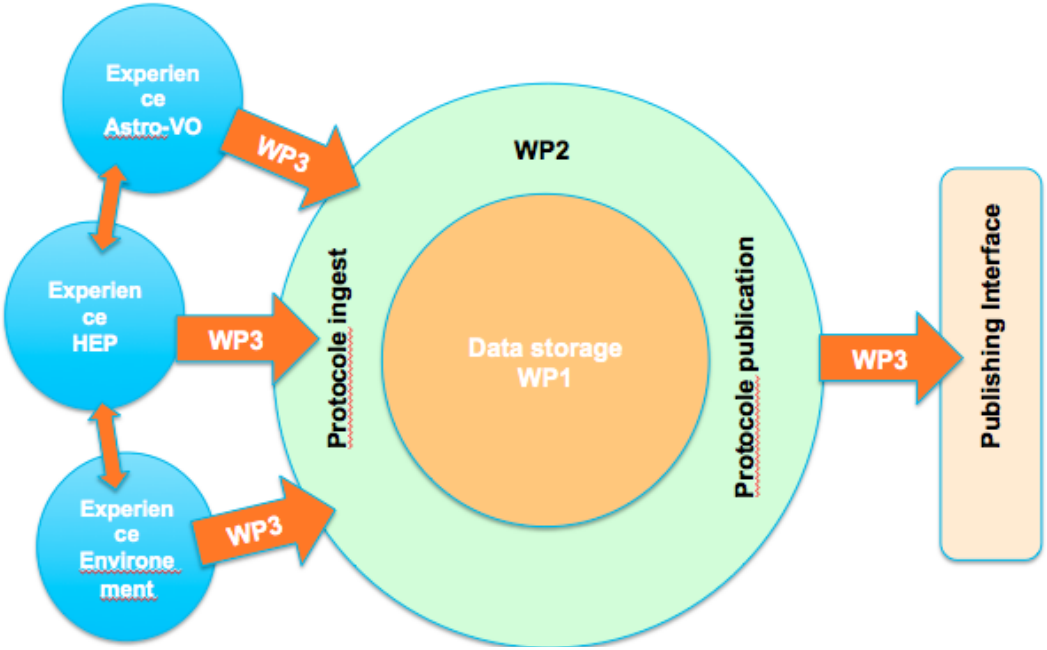
LabEx UnivEarths project at APC / François Arago Centre:

- potential of the cloud versus classical data processing and storage opportunities
- test processing on Francois Arago Centre cluster, compared with Cloud StratusLab

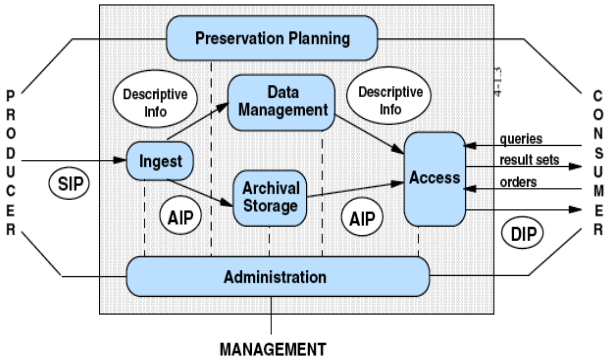
Schematic description of the cloud StratusLab, which is a European public cloud project IaaS which started in 2010.



PREDON: Concept demonstrator



OAIS



- But : « forcer » les frontières entre les disciplines, par exemple:
  - essayer des formats astrophysique (VOT) et des outils de visualisation (Tulip) sur des données HEP
  - Stocker des données complexes et très « custom » dans un projet de sauvegarde de données généraliste (ISAAC)

Interface données HEP – ISAAC (CINES)

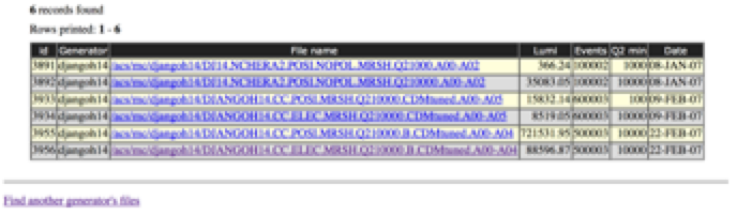
PREDON 01/2013

Production Monte Carlo (données simulées) de la Collaboration H1/DESY : Interface vers le projet ISAAC-CINES

La production Monte Carlo s'appuie sur des logiciel nommé "générateurs", l'output de ces logiciels est encore procéssé en « simulation » et ensuite « reconstruction ».

Pour stocker (et ensuite être capable de trouver) ces données l'interface est la suivante :

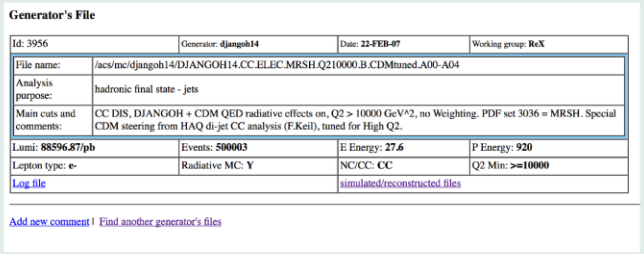
Si je choisi par exemple le generateur « django 14 » je trouve une liste de « productions » qui correspondent à cette requette. Les productions sont l'unité identifié par un « ID » de generateur unique



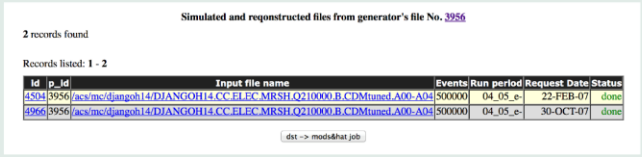
Si je clique sur la première production , je retrouve des meta-donnes, en particulier, le nom de l'ensembkle des fichiers, ainsi que un pointeur sur la « log file » qui contient

Niveau information package

- **Proposition : l'unité archivée est la production**
  - Les métadonnées reprennent les informations de ce niveau



  - Et pour chaque fichier la composant….



CINES Centre Informatique National de l'Enseignement Supérieur

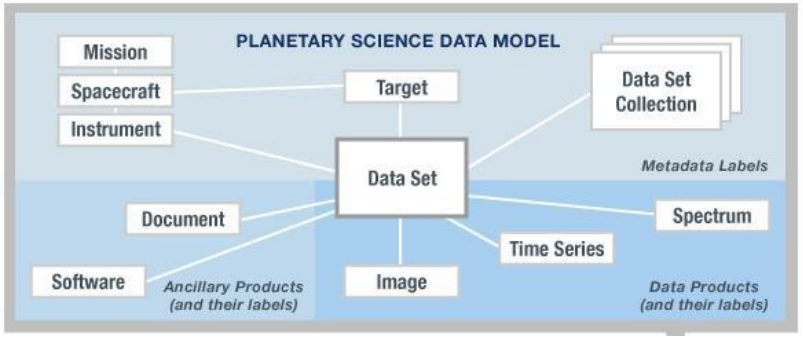
Hands on ISAAC – workshop PREDON 14 nov 2013

ICDE conference Chicago March2014

http://lipade.math-info.univ-paris5.fr/lops/?page_id=96

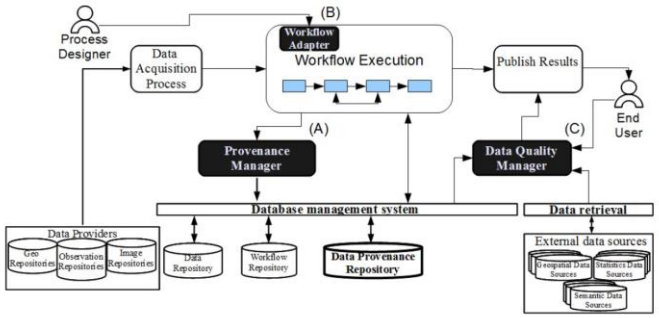
- Workshop LoPS
  - organisé par PREDON
  - Journée dédiée à la préservation des données scientifiques
  - Agenda:

PDS- A Model-Driven Planetary Science Data Architecture for Long-Term Preservation
John Hughes (Jet Propulsion Laboratory, NASA)



A provenance-based approach to manage long term preservation of scientific data.
Claudia Medeiros (University of Campinas)

Quality aware workflows



Invest in capturing and maintaining data in **well-annotated, accessible, structured data repositories**

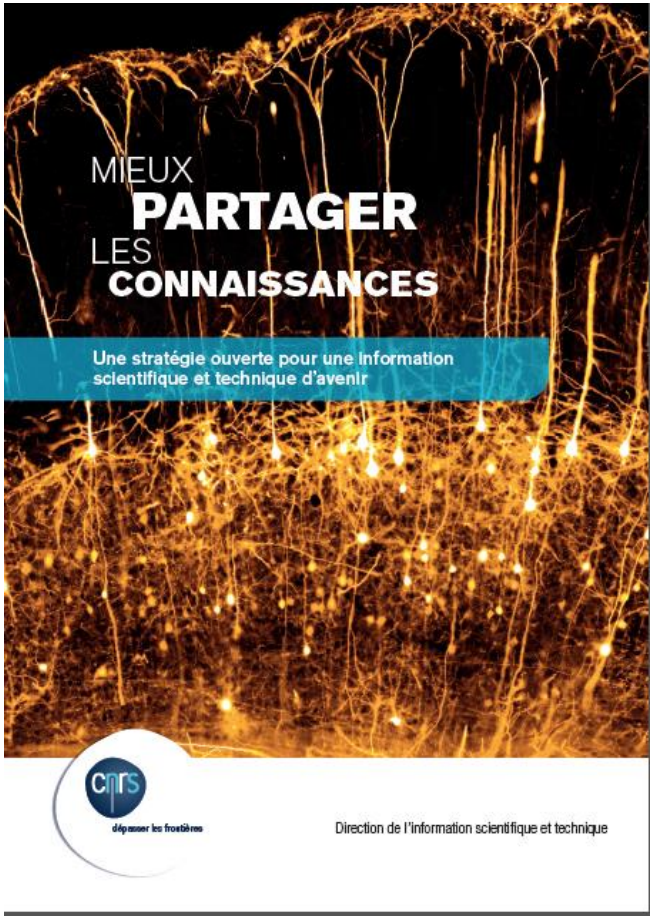
Computer Scientists, Statisticians/Data Scientists, Domain Experts (Scientists) must **systematize the analysis of massive data**

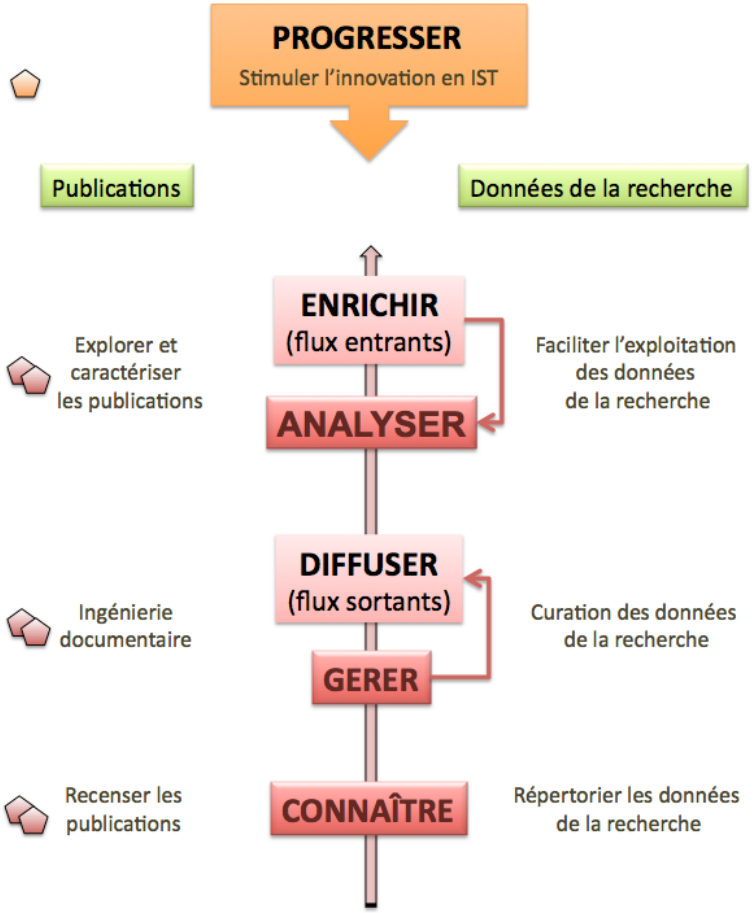
Develop **computing infrastructures** for sharing and analyzing highly distributed, heterogeneous data

Sustainability in both the data and the software infrastructures are critical

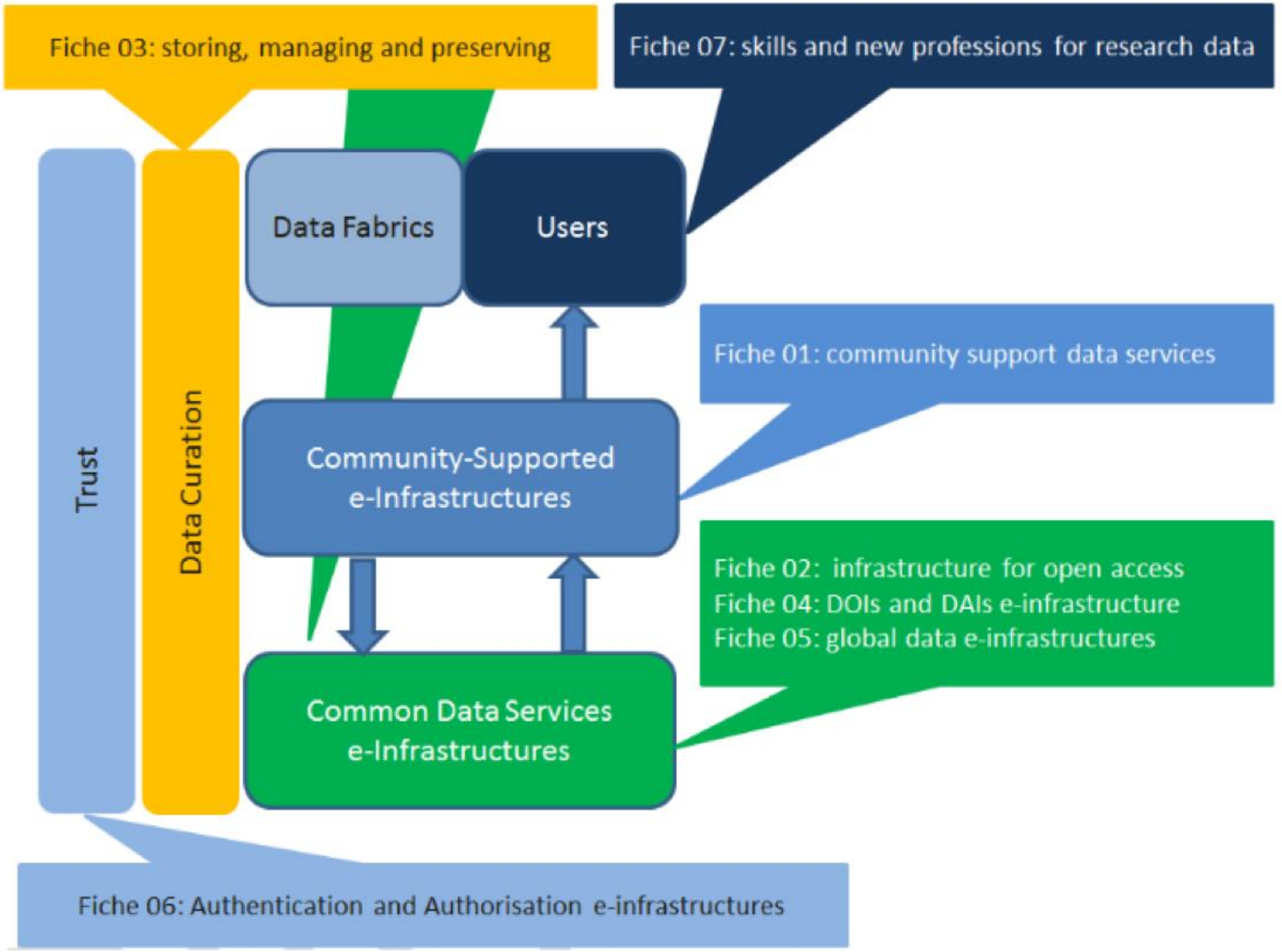
Strategie nationale IST

- PAP3: Analyser et valoriser l'information





Opportunités H2020



Agenda

Wednesday, 5 November 2014	
09:00	**Case for Preservation** (until 10:30)
09:10	The golden mine of the future: scientific data preservation - Cristinel Diaconu (CPPM, Aix-Marseille Université, CNRS/IN2P3 (FR))
09:40	Towards a public analysis database for LHC new physics searches - Sabine Kraml (Centre National de la Recherche Scientifique (FR))
11:00	**Methodology for Data Preservation** (until 13:05)
11:00	Le pôle de recherche sur la conservation des données numériques (GIS-SPADON) - Pollack Jean-Dominique (Institut Jean le Rond d'Alembert, Université Pierre et Marie Curie/CNRS)
11:35	Le projet pluri-disciplinaire IDV (Imagerie du vivant) de l'Université Sorbonne Paris Cité et quelques réflexions / méthodologies liées aux données. - Christophe Cérin (urn:Google)
12:10	Data preservation methodology at CNES/ CCSDS - Danielle Boucon (CNES)
14:00	**Methodology for Data Preservation** (until 16:00)
14:00	Préservation de données dans le contexte IndexMed - Romain David (IMBE)
14:45	Perenisation des donnees au CDS (Centre de données astronomiques de Strasbourg) - Gilles Landais (Centre de Données Astronomiques de Strasbourg)
15:30	Préservation de données en imagérie médicale - Pierre Bourdoncle
16:30	**Technologies for Data Preservation** (until 17:45)
16:30	Cloud technology for algorithms preservation - Cécile Cavet (APC/ Univ. Paris 7) summary
17:00	EUDAT - STEPHANE COUTIN (C)
17:45	**Discussion** (until 18:30)
20:00	**Social Dinner** (until 23:00) ()

Objectifs du workshop

- Tour des projets au sein de PREDON
- Récevoir des points de vue d'autres disciplines
  - Nouvelles approches: documentation, juridique, économique
- Connexion aux projets similaires en France
- Document: **« Scientific Data Preservation 2015 »**
  - Contributions suite aux workshops PREDON
  - Contributions invitées: propositions?

Conclusions

- 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

  - Un GRAND merci aux organisateurs!

BACKUP

Groupe d'études PREDON

2013 2012 Prop.

- **> IN2P3**
  - Cristinel Diaconu, Dirk Hofmann, Angélique Pèpe, Magali Damoiseaux, D. Christofol (CPPM, Marseille)
  - Sabine Kraml (LPSC, Grenoble)
  - Giovanni Lamanna (LAPP, Annecy)
  - Volker Beckmann (APC, Centre Francois Arago, Paris 7)
- **> CCIN2P3**
  - Ghita Rahal, Jean-Yves Nief (CC-IN2P3)
- **> INSU**
  - Christian Surace (LAM/OAMP Cesam, Marseille)
- **> INS2I**
  - Mustapha Lebbah (LIPN, Paris 13)
  - Salima Benbernou (LIPADE, Paris 5)
  - Anne Laurent, Sophie Nicoud (LIRMM, Montpellier)
- **> CINES**
  - Stéphane Coutin, Marion Massol (CINES, Montpellier)
- **> IRD**
  - Thérèse Libourel, Yuan Lin (Espace DEV)

Nouveau contacts en 2013/2014 suite aux workshops:

Daniel Chateigner, CRISMAT/ENSICAEN, données **cristallographie**

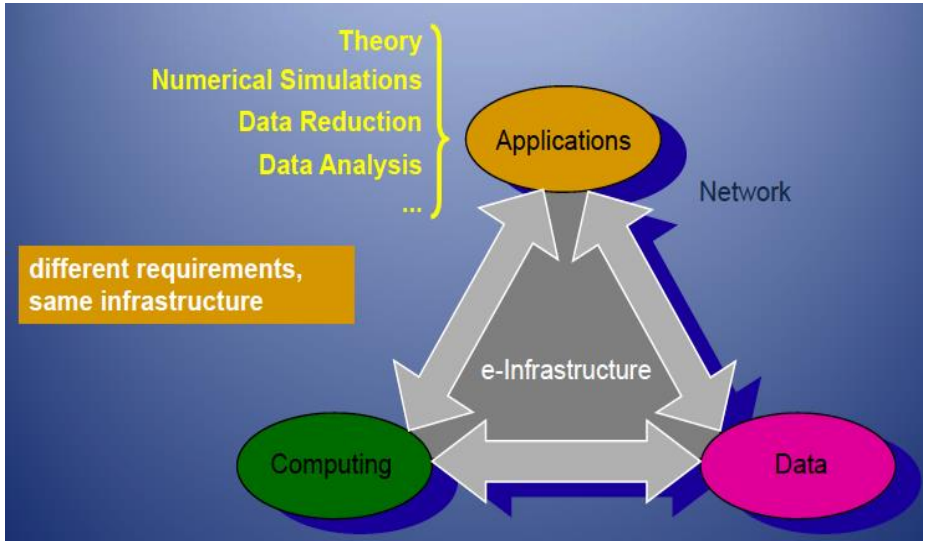
Marc Schaming, Institut de **Physique du Globe** (CNRS/UNISTRA), IPG Strasbourg

Catherine Boisson de l'Observatoire de Meudon / LUTH/**INSU** CTA

Danièle Boucon, expert en préservation de données **CNES**

Jean-Dominique Pollack, LAM, UPMC, GIS **SPADON**

Virtual Observatories in Astrophysics



- Data Archives Inter-operable
- Work on standards and access to
  - Data, simulation, mining techniques
- International, multi-experiment
- Agregated Person-power: about 100FTE

Préservation des connaissances

- Le stockage des données à long terme demande une organisation rigoureuse
- Le vrai challenge technique est la préservation des connaissances « meta-digitales »

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 plots specific to each experiment. These often degrade or disappear with time, says Cristinel Diaconu of the Marseilles Centre for Particle Physics in France

nature

DPHEP : définition des niveaux de préservation

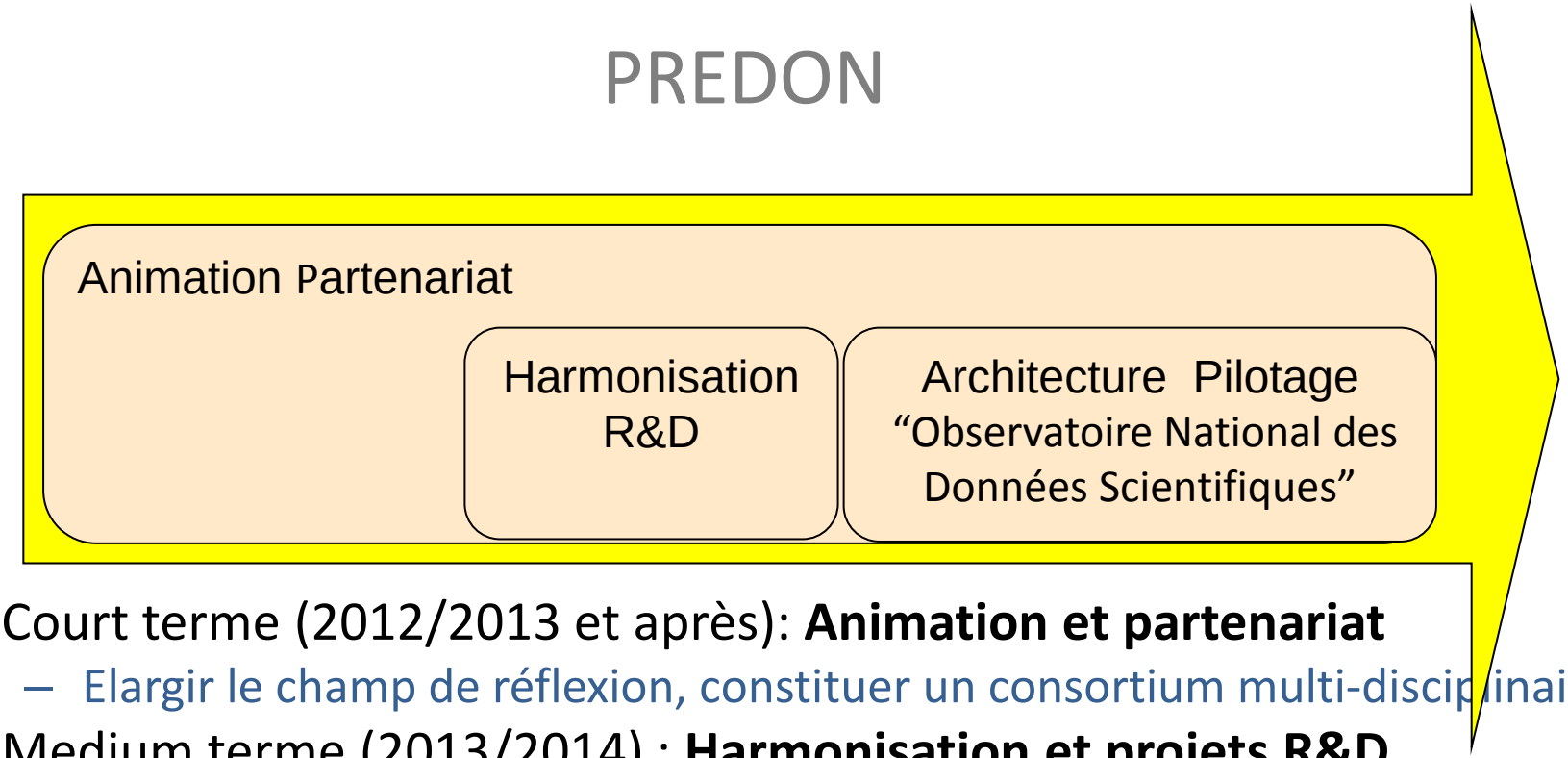
- En progression de la complexité et les couts

Preservation Model		Use Case	
1	Provide additional documentation	Publication related info search	**Documentation**
2	Preserve the data in a simplified format	Outreach, simple training analyses	**Outreach**
3	Preserve the analysis level software and data format	Full scientific analysis, based on the existing reconstruction	**Technical Preservation Projects**
4	Preserve the reconstruction and simulation software as well as the basic level data	Retain the full potential of the experimental data	

MASTODONS

- Stockage et gestion de données (par exemple, dans le Cloud), sécurité, confidentialité.
- Calcul intensif sur des grands volumes de données, parallélisme dirigé par les données.
- Visualisation de grandes masses de données.
- Extraction de connaissances, datamining et apprentissage.
- Qualité des données, confidentialité et sécurité des données.
- Problèmes de propriété, de droit d'usage, droit à l'oubli.
- **Préservation/archivage des données pour les générations futures.**
  - **PREDON (PREservation des DONnees)**

PREDON



- Court terme (2012/2013 et après): **Animation et partenariat**
  - Elargir le champ de réflexion, constituer un consortium multi-disciplinaire
- Medium terme (2013/2014) : **Harmonisation et projets R&D**
  - Communication: exchanges and workshops
  - Livre blanc sur la préservation et la mise à disposition des données scientifiques dans un contexte multi-disciplinaire
  - Démonstrateur accès et préservation de données scientifiques complexes
- Long term (2015/2016) **Architecture et pilotage**
  - "Observatoire National des Données Scientifiques"
    - Coalition de grands centres de données et projets multi-disciplinaires
    - Support et suivi des lots de données scientifiques : accès et préservation

Workshop on Data Preservation at ICDE 2014

LOPS@ICDE

WORKSHOP ON LONG TERM PRESERVATION FOR BIG SCIENTIFIC DATA

- Coordonnateurs workshop: S.Benbernou, C. Diaconu
- http://lipade.math-info.univ-paris5.fr/lops/
- LOPS will be held in conjunction with the 30th IEEE International Conference on Data Engineering. Chicago, IL, USA. March 31-April 4, 2014.

PREDON 2014

- Organisation Workshop LOPS@ ICDE2014
- Continuation et initiation de nouveaux mini-projets et démonstrateurs pour des cas spécifiques de préservation de données (stages)
  - HEP-Data @ ISAAC
  - Formats de données transdisciplinaires
- Réunions du groupe de travail : nouveau contacts, séminaires
  - Extensions possibles à d'autres domaines (bio, IST, économie)
  - Aborder des questions communes (cout, persistence, open access, éducation, outreach etc.)
- Organisation d'un Atelier sur la préservation des données scientifiques et en relation avec la thématique « Big Data »
  - Publication PREDON: 2015
- Participation aux groupes de travail au niveau international et aux projets et consortia en cours de constitution pour des programmes de financement H2020.

Site web PREDON

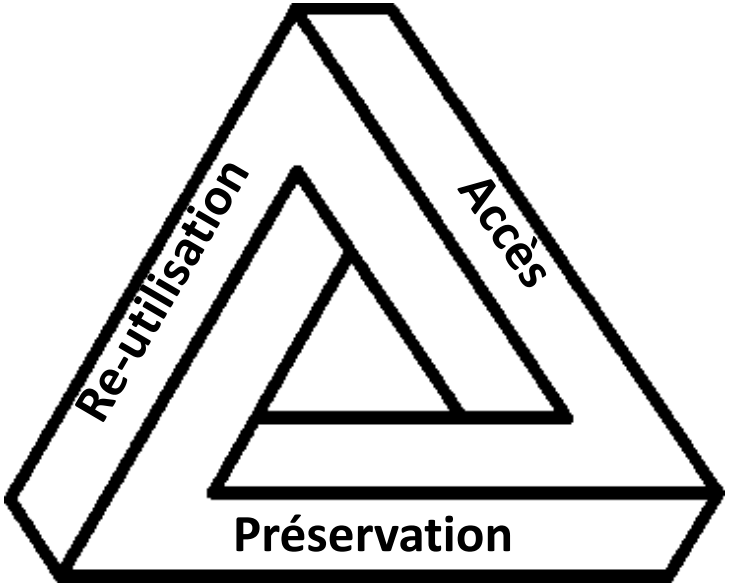
http://predon.org

Summary of information from the (pre-LHC) experiments

	BaBar	H1	ZEUS	HERMES	Belle	BESIII	CDF	DØ
End of data taking	07.04.08	30.06.07	30.06.07	30.06.07	30.06.10	2017	30.09.11	30.09.11
Type of data to be preserved	RAW data Sim/rec level Data skims in ROOT	RAW data Sim/rec level Analysis level ROOT data	Flat ROOT based ntuples	RAW data Sim/rec level Analysis level ROOT data	RAW data Sim/rec level	RAW data Sim/rec level ROOT data	RAW data Rec. level ROOT files (data+MC)	Raw data Rec. level ROOT files (data+MC)
Data Volume	2 PB	0.5 PB	0.2 PB	0.5 PB	4 PB	6 PB	9 PB	8.5 PB
Desired longevity of long term analysis	Unlimited	At least 10 years	At least 20 years	5-10 years	5 years	15 years	Unlimited	10 years
Current operating system	SL/RHEL3 SL/RHEL 5	SL5	SL5	SL3 SL5	SL5/RHEL5	SL5	SL5 SL6	SL5
Languages	C++ Java Python	C C++ Fortran Python	C++	C C++ Fortran Python	C C++ Fortran	C++	C C++ Python	C++
Simulation	GEANT 4	GEANT 3	GEANT 3	GEANT 3	GEANT 3	GEANT 4	GEANT 3	GEANT 3
External dependencies	ACE CERNLIB CLHEP CMLOG Flex GNU Bison MySQL Oracle ROOT TCL XRootD	CERNLIB FastJet NeuroBayes Oracle ROOT	ROOT	ADAMO CERNLIB ROOT	Boost CERNLIB NeuroBayes PostgresQL ROOT	CASTPR CERNLIB CLHEP HepMC ROOT	CERNLIB NeuroBayes Oracle ROOT	Oracle ROOT

Longévité recherchée: > 10 ans

Préservation, réutilisation, libre-accès



→ La préservation suppose la mise à disposition en accès libre

→ Maximiser le bénéfice

← Le libre-accès facilite la préservation à long terme

← Elargir la communauté, multiplier les connaissances

Generic arguments

- Task forces already in place to address this issue in a generic way (standards)
  - e.g. Blue Ribbon, APA, DPC, eSciDir, …

http://www.alliancepermanentaccess.eu
http://brtf.sdsc.edu

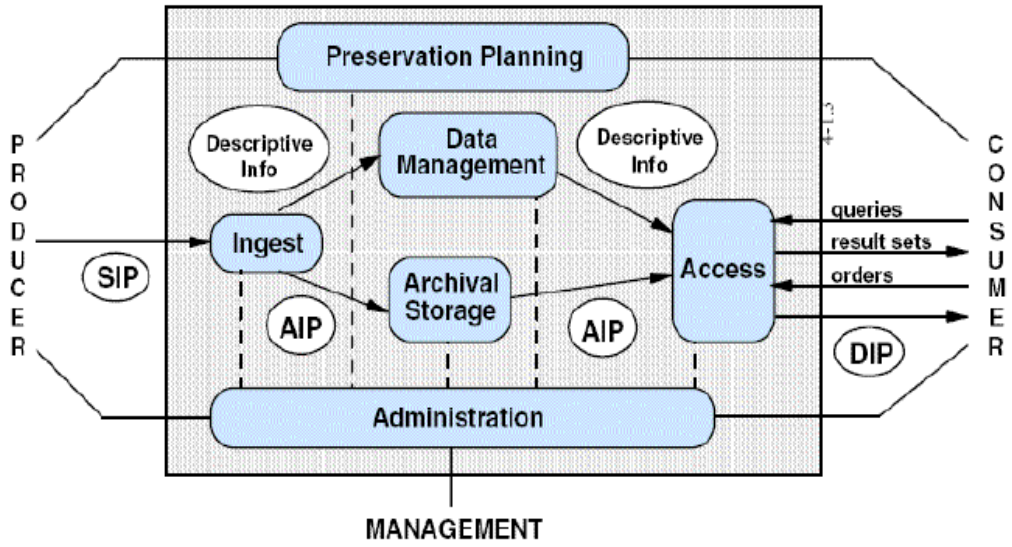


FIGURE 2.1: **The OAIS Reference Model**
http://public.ccsds.org/publications/archive/650x0b1.pdf, Page 4-1.
Source: Consultative Committee for Space Data Systems January 2002.

- Scientific Data is a major component of the ongoing efforts (complexity)

Documentation

- Une tache considérable

- Non-digital: Cataloguing, organisation, scanning or photographing of appropriate of papers, notes, drawings, talks from pre-web days, detector schematics, blueprints, logbooks, ...
  - *Virtual Archives* established by the experiments
- Digital: Old online shift tools, detector configuration files, electronic logbooks, detailed run information, web content from out-dated servers with dead links, various wikis, meetings, talks, ...
  - Replacement of old web servers by VMs, hosted by the computer centres
  - Replacement of old pages to newer technologies such as wikis (use of (T)wikis much more prevalent in the LHC era)
  - Use of external services for hosting collaboration material

PREDON: Challenges

- **Scientific Potential Challenge**: these data sets contain unexploited information, which may give rise to highly useful for joint, multi-disciplinary project.
- **Complexity Challenge**: the data collected by the experimental devices considered in the project is unique and encodes a large typology, well beyond the regular, well-structured data produced in large quantities in the industrial world**.**
- **Technological et methodological challenge.** The installation of procedures, workflows, algorithms for long term data preservation, as well as the definition of suitable technological frameworks constitute novel investigation domains.