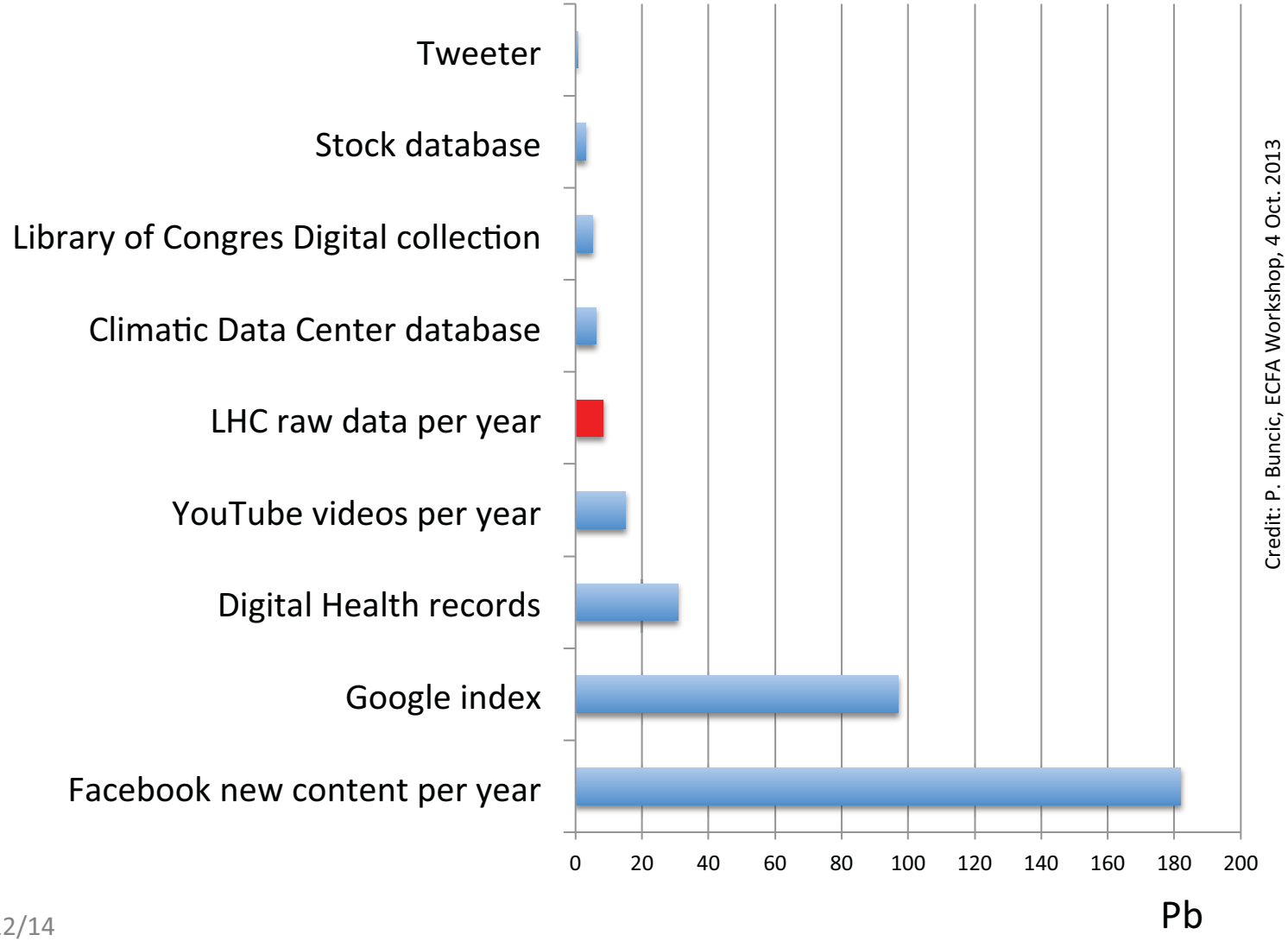
La préservation des données scientifiques

une mine d'or pour la science de demain

C. Diaconu
Centre de Physique des Particules de Marseille
CNRS et Aix Marseille Université (AMU)

predon.org
dphep.org

Données digitales explosent (les données scientifiques aussi)



Les données digitales sont fragiles

- En plus, la capacité de stockage est physiquement dépassée depuis longtemps

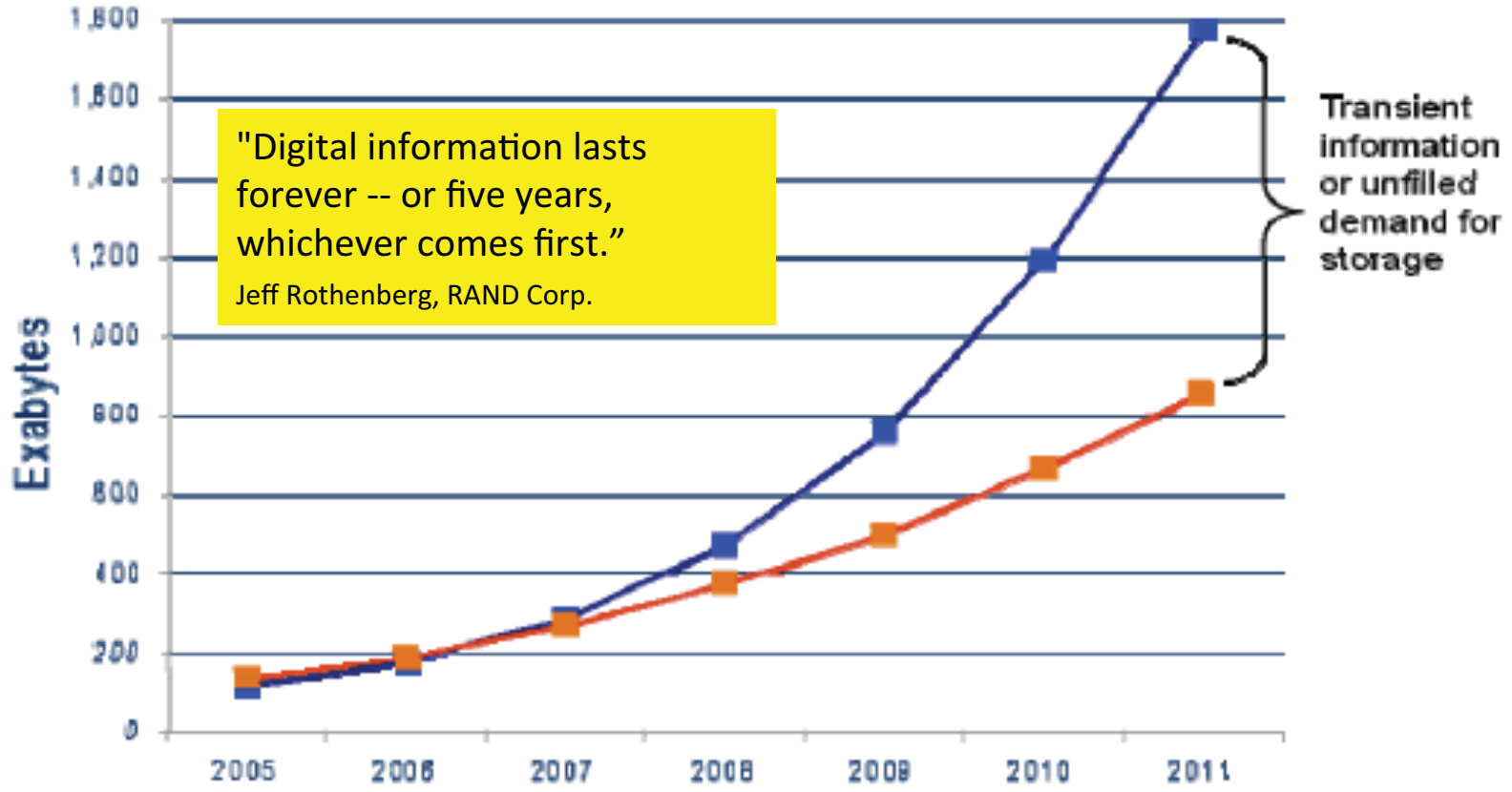
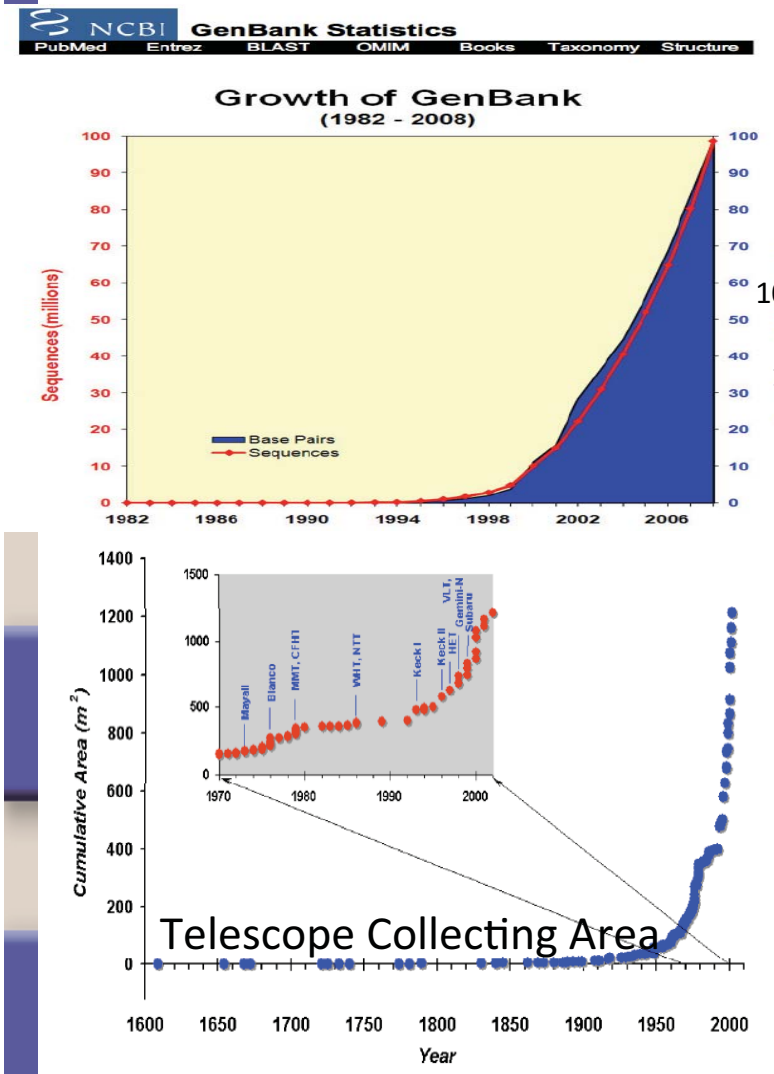
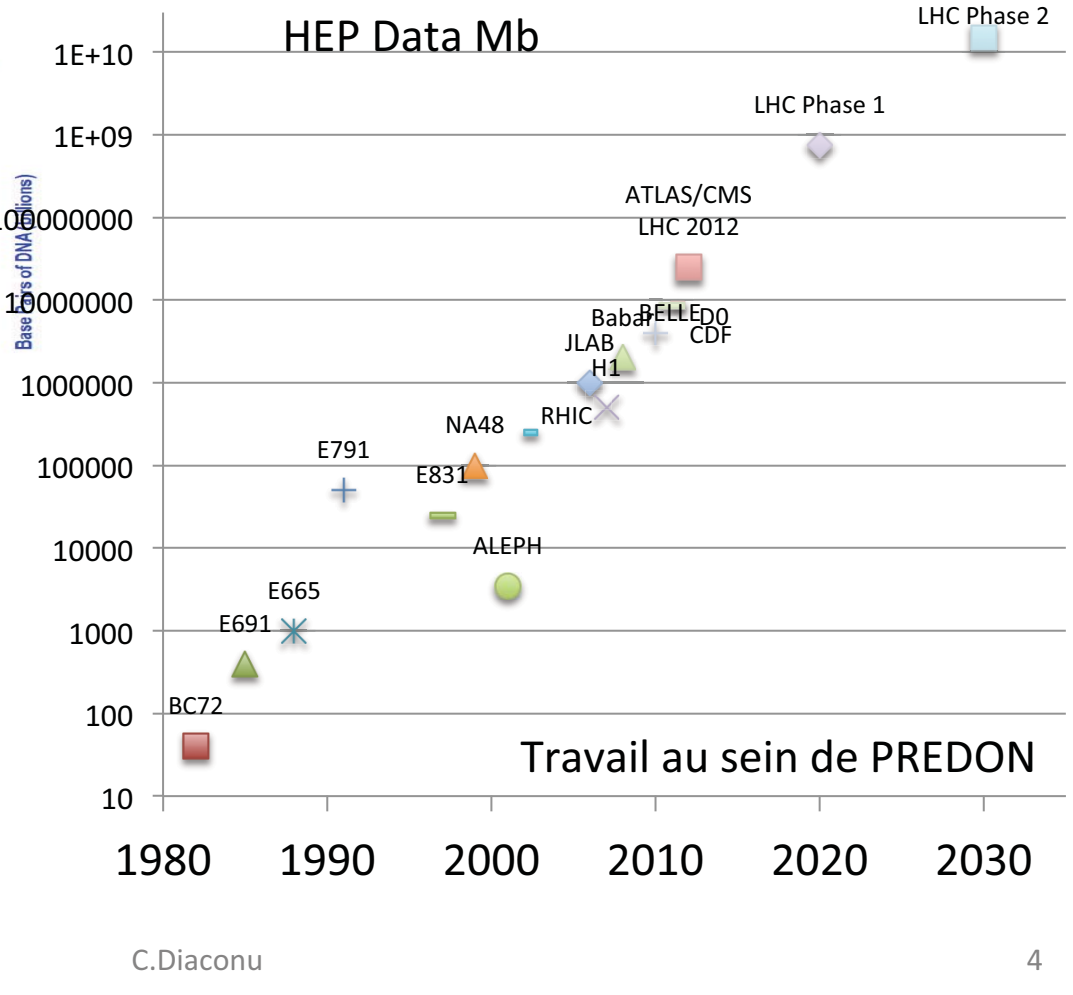


FIGURE 1.3: **Information and Storage**
Source: J. Gantz January 2008 (revised). Used with permission.

« Big Scientific Data »

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





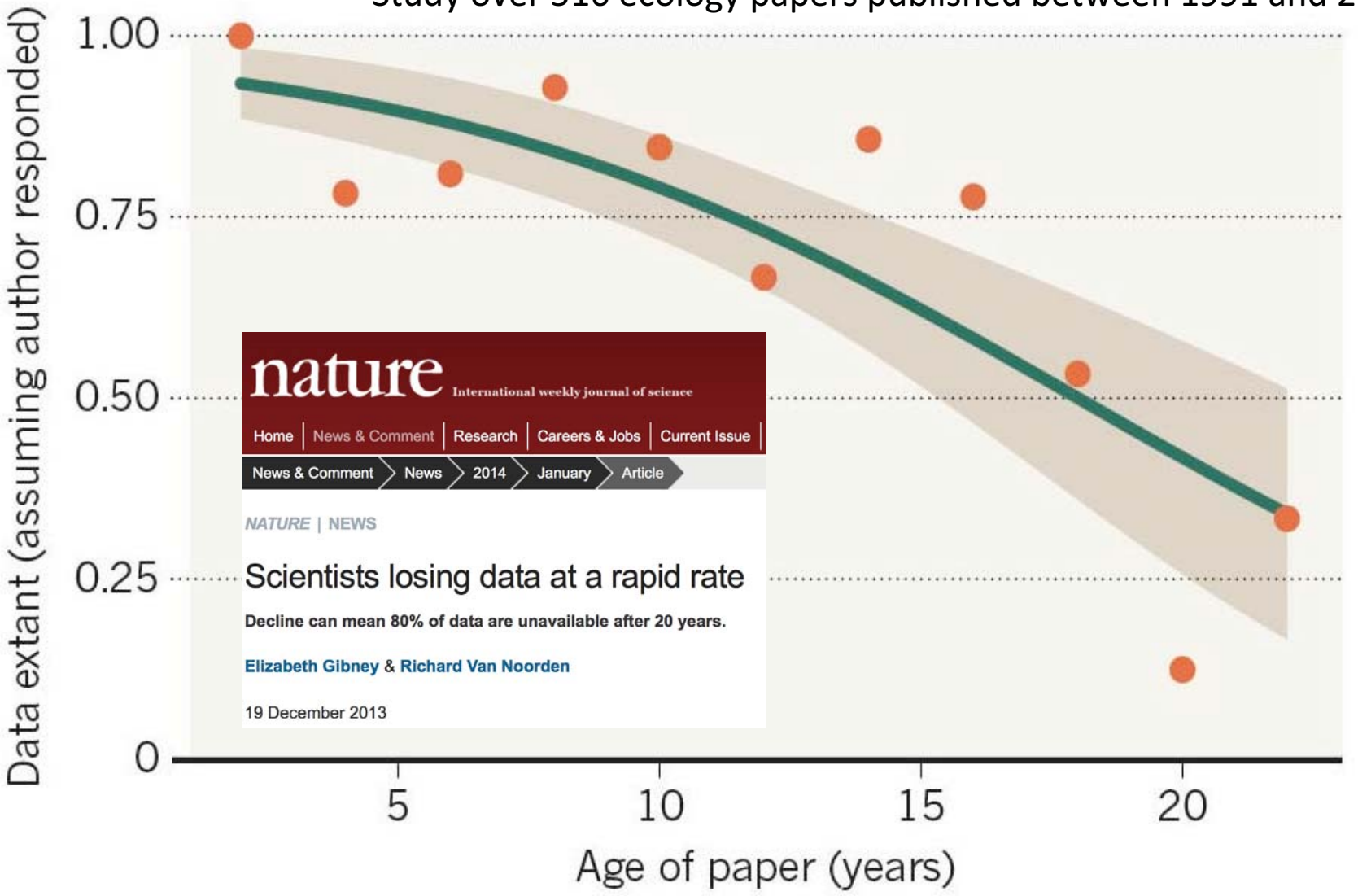
Est-ce que les données scientifiques sont spéciales (« big » à part)?

- 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
- PRESERVATION!

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

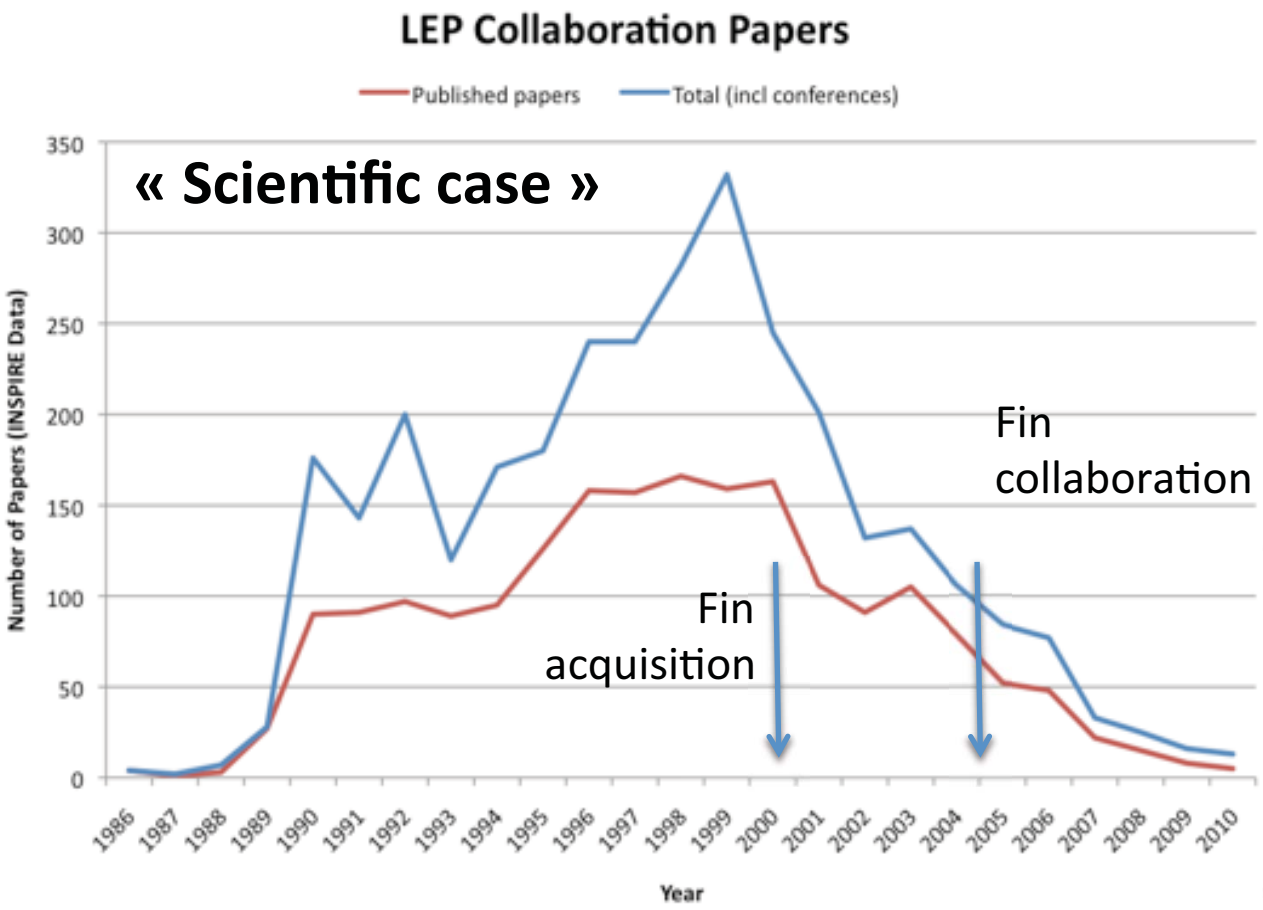
Scientists losing data at a rapid rate

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

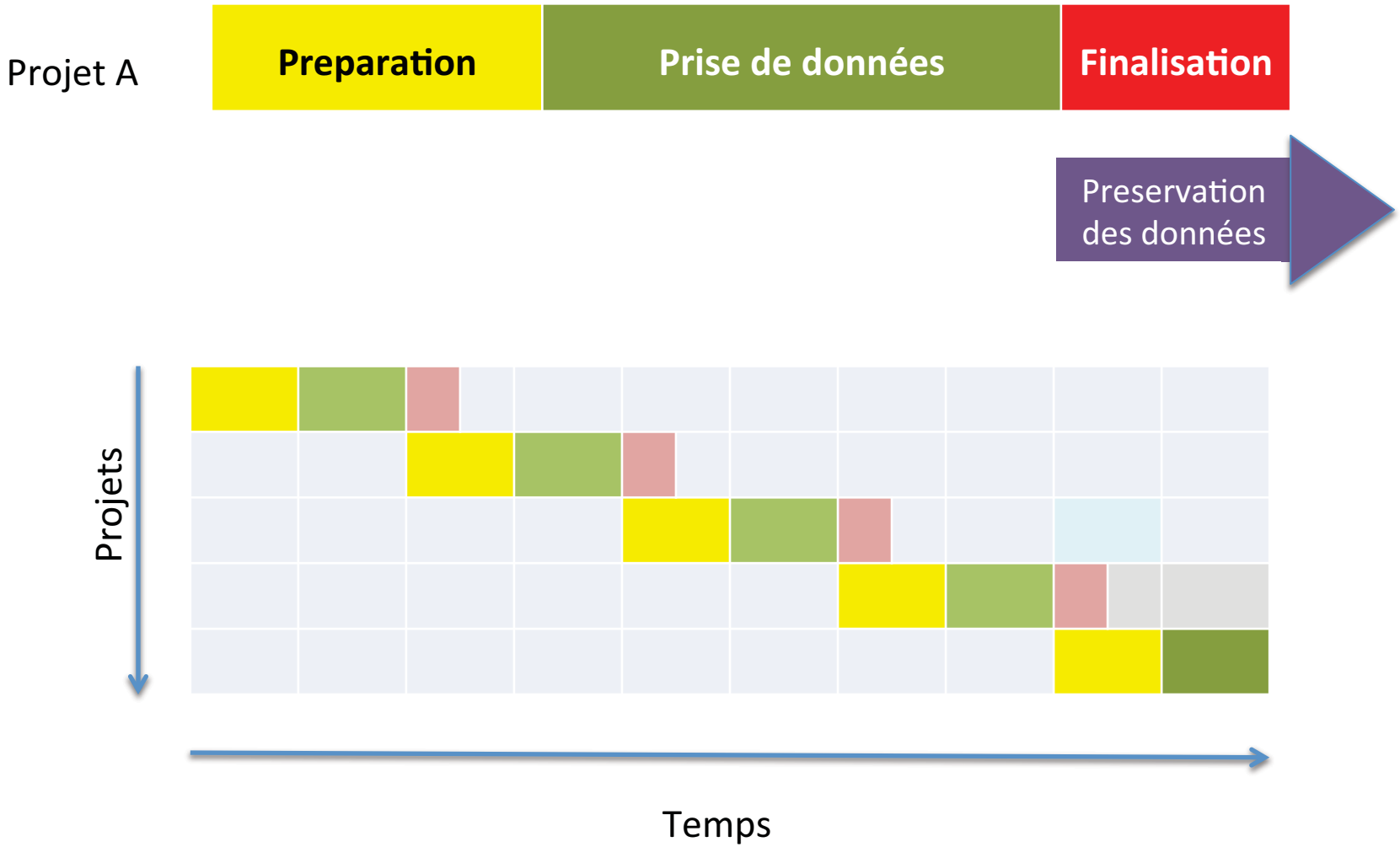
Elizabeth Gibney & Richard Van Noorden

19 December 2013

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



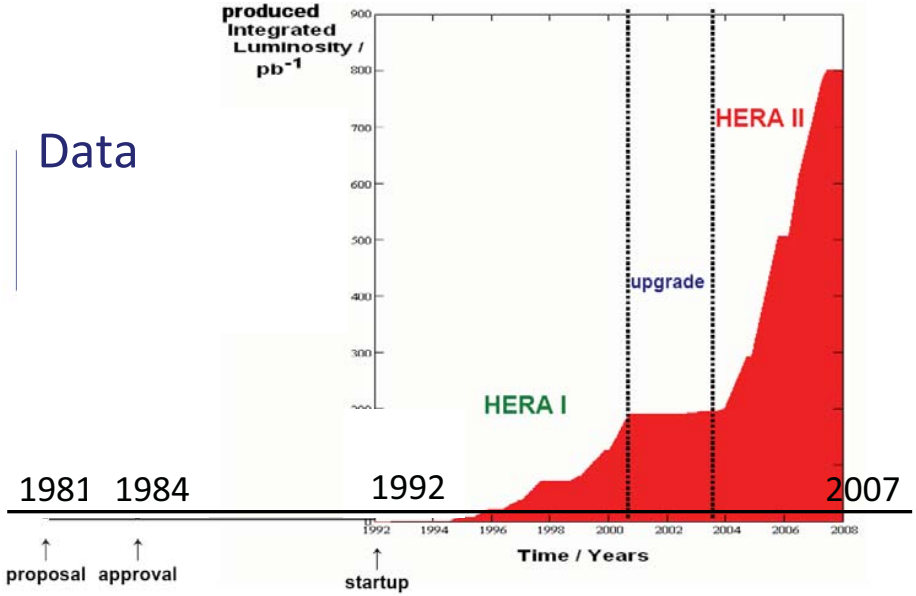
Quand faut-il commencer à préserver?

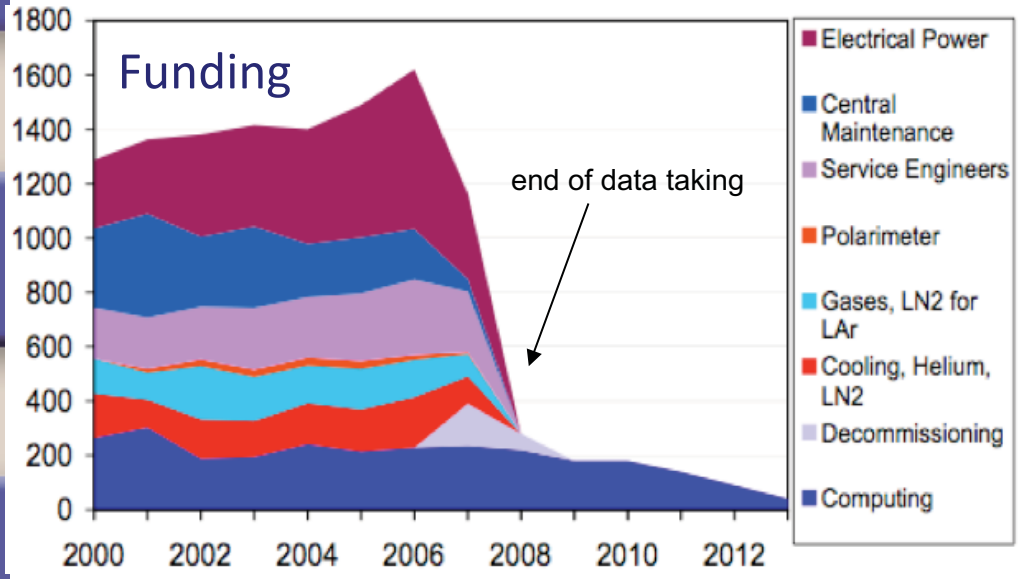


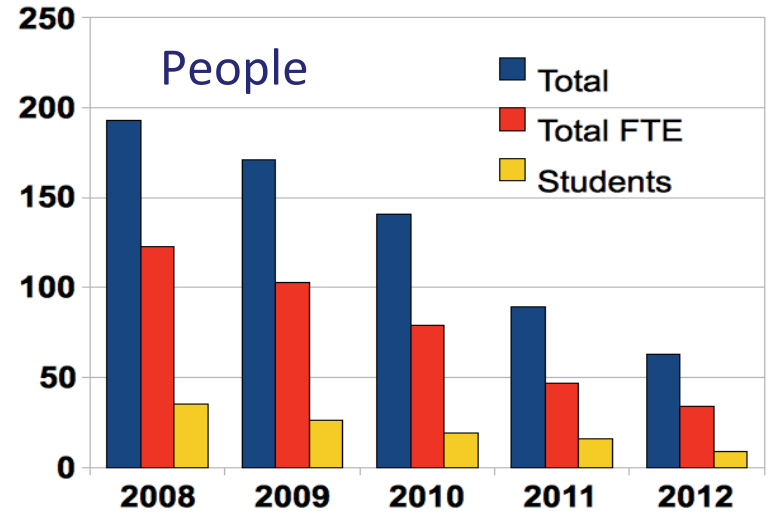
Le contexte de la préservation de données

- 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

Data

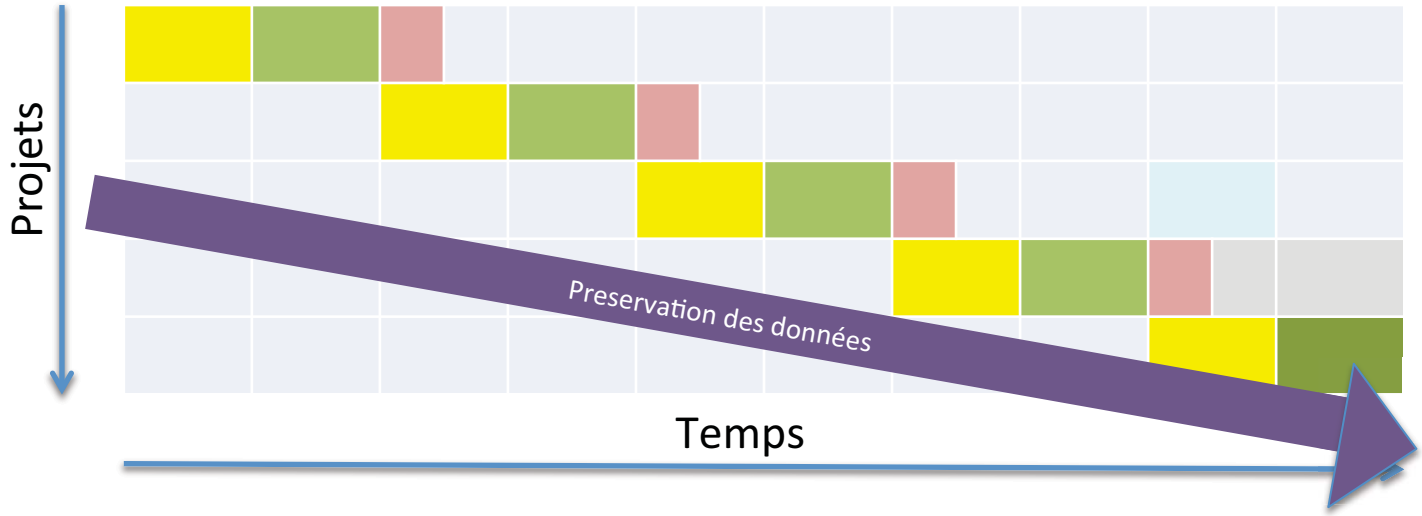






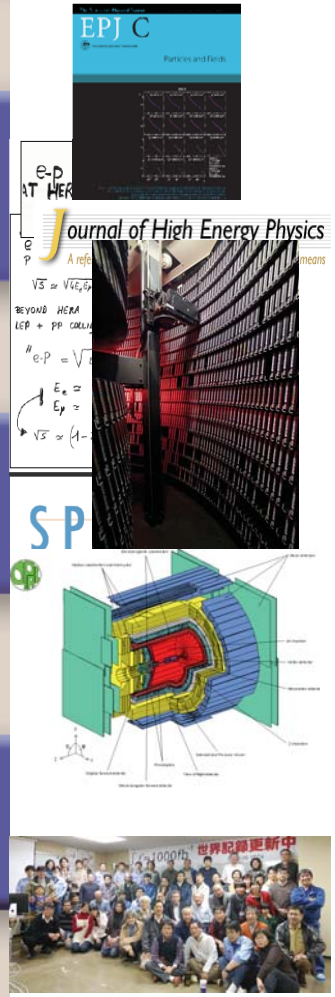
Quand faut-il commencer à préserver?

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



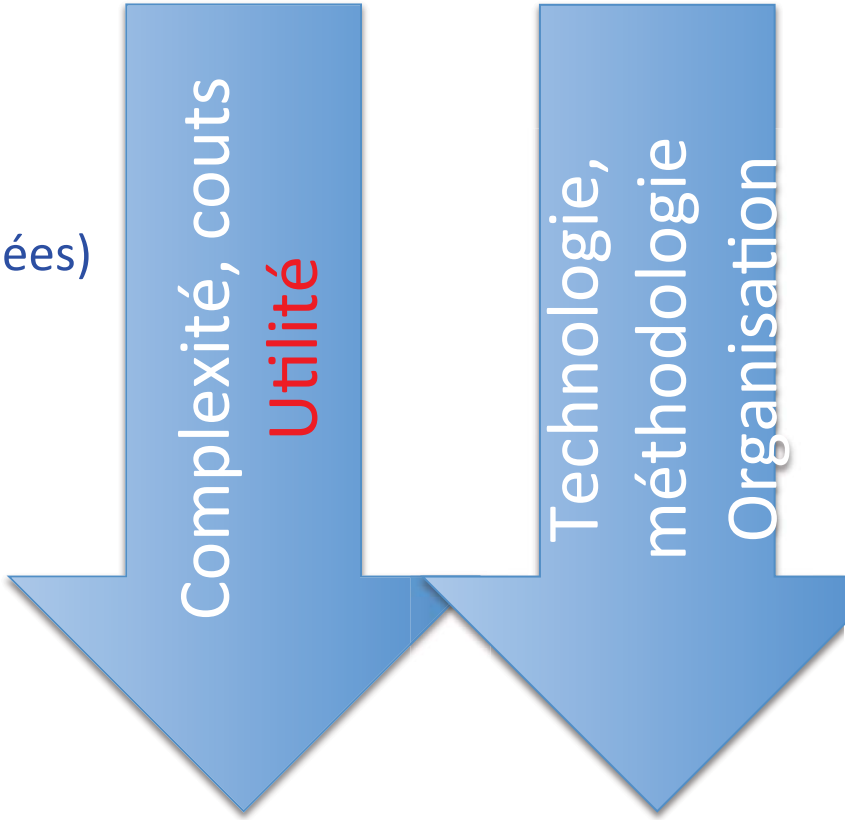
Programme cohérent de la préservation de données

Données Scientifiques



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

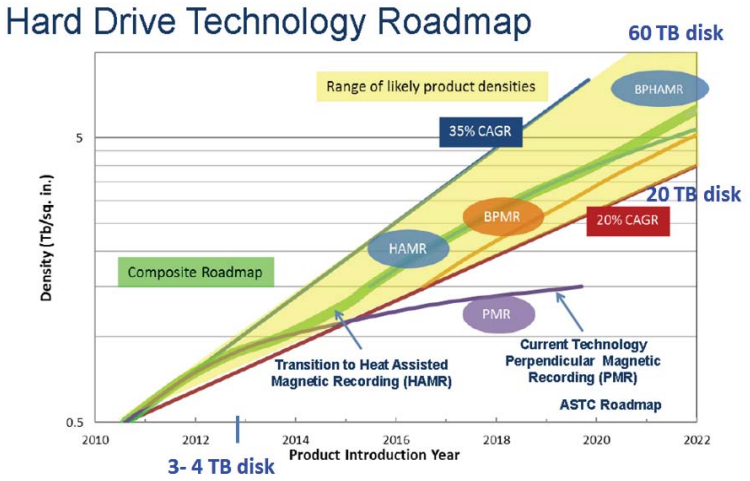
….more…



Quel modèle de préservation pour les données scientifiques?

Challenges

- Sauver le « octets »
  - Centres de données (distribués)
  - Couts?
    - 2x taille initiale (1+1/2+1/4+….)
- Organisation
  - Indexation, metadata, standards,…(OAIS)
  - Documentation, connaissances
  - Collaborations à long terme
- Software: complexe, fragile



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

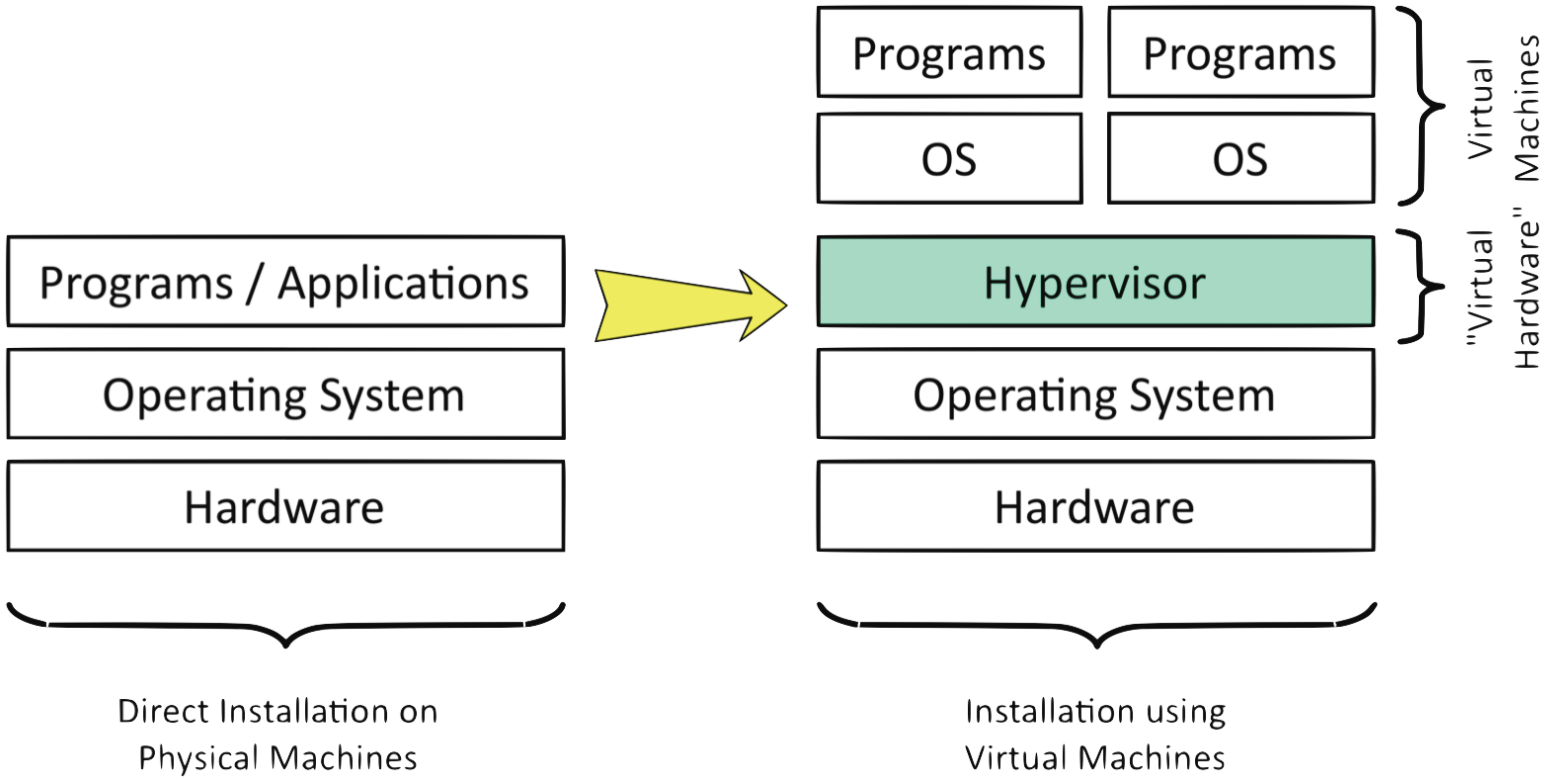
Generic models for Data Preservation

- 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

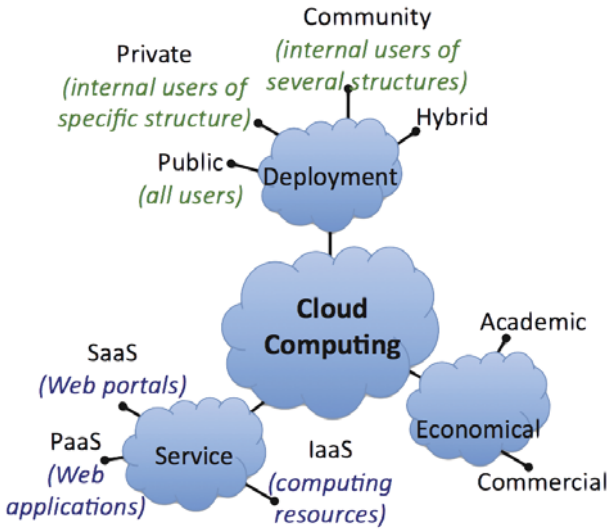
Préservation de données et virtualisation

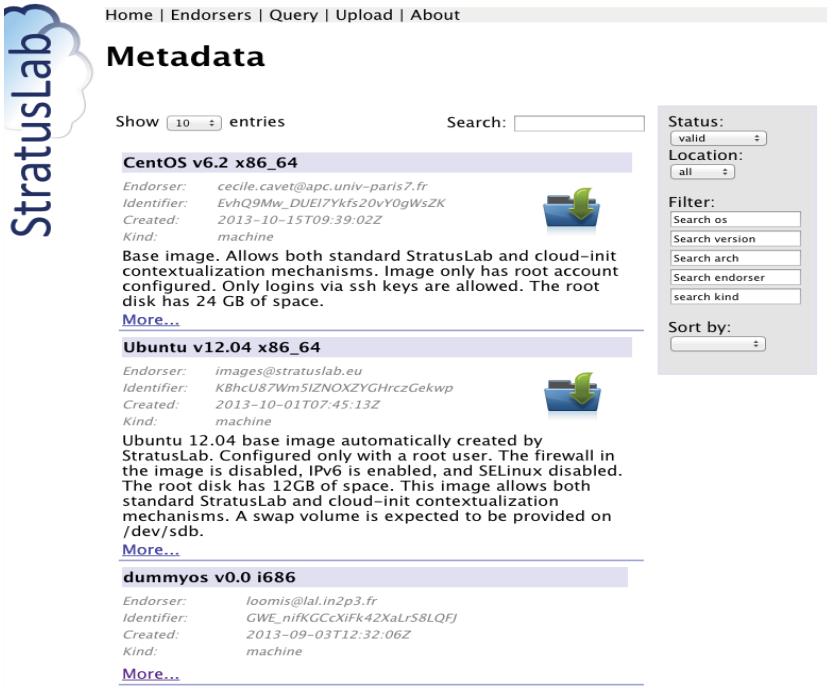


C. Loomis http://indico2.lal.in2p3.fr/indico/getFile.py/access?contribId=0&resId=0&materialId=slides&confId=1897

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

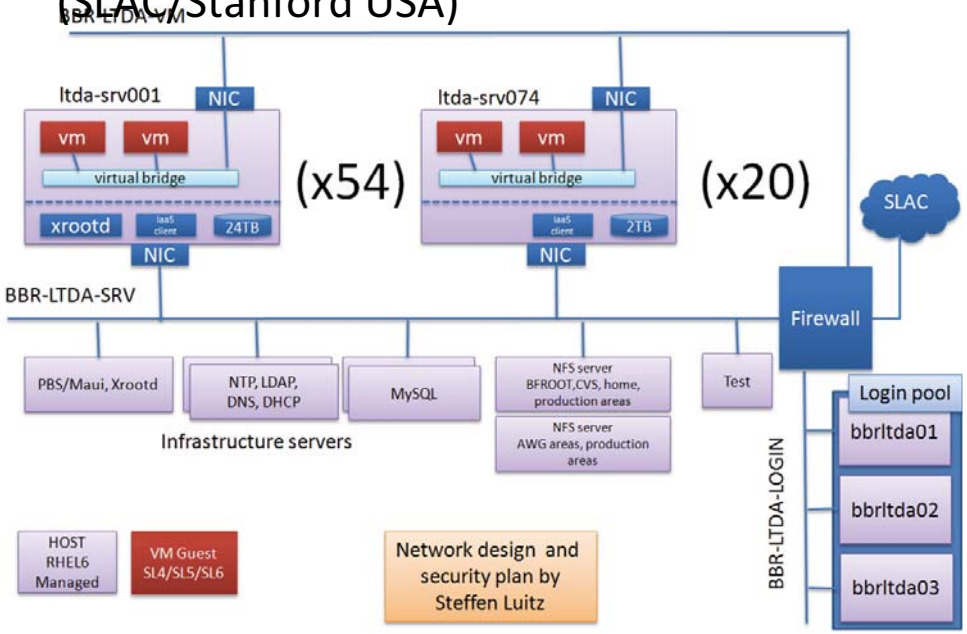
Disk images have **6 months of validity**
OS update/upgrade for security.

Virtualisation/cloud need to be tuned for long term

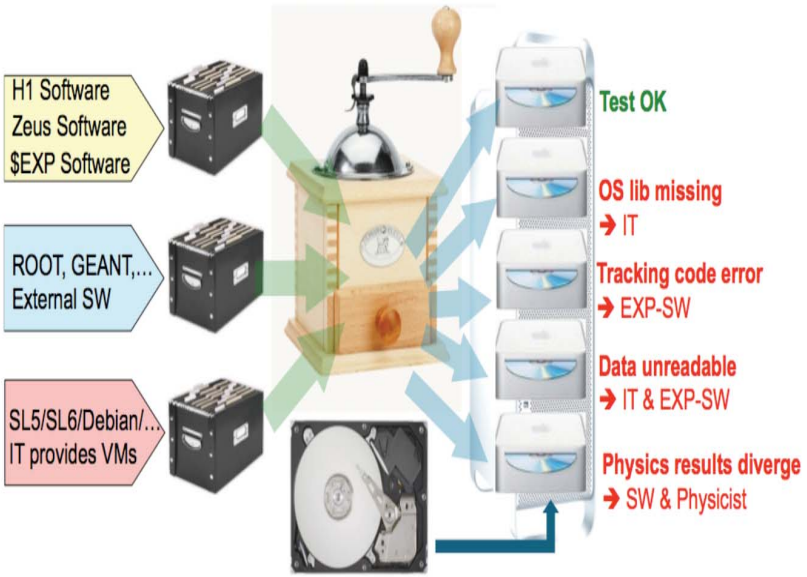
Physique des Particules

dphep.org

Préservation d'un système d'accès et calcul à des données complexes
Basé sur une **ferme virtuelle** (SLAC/Stanford USA)



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



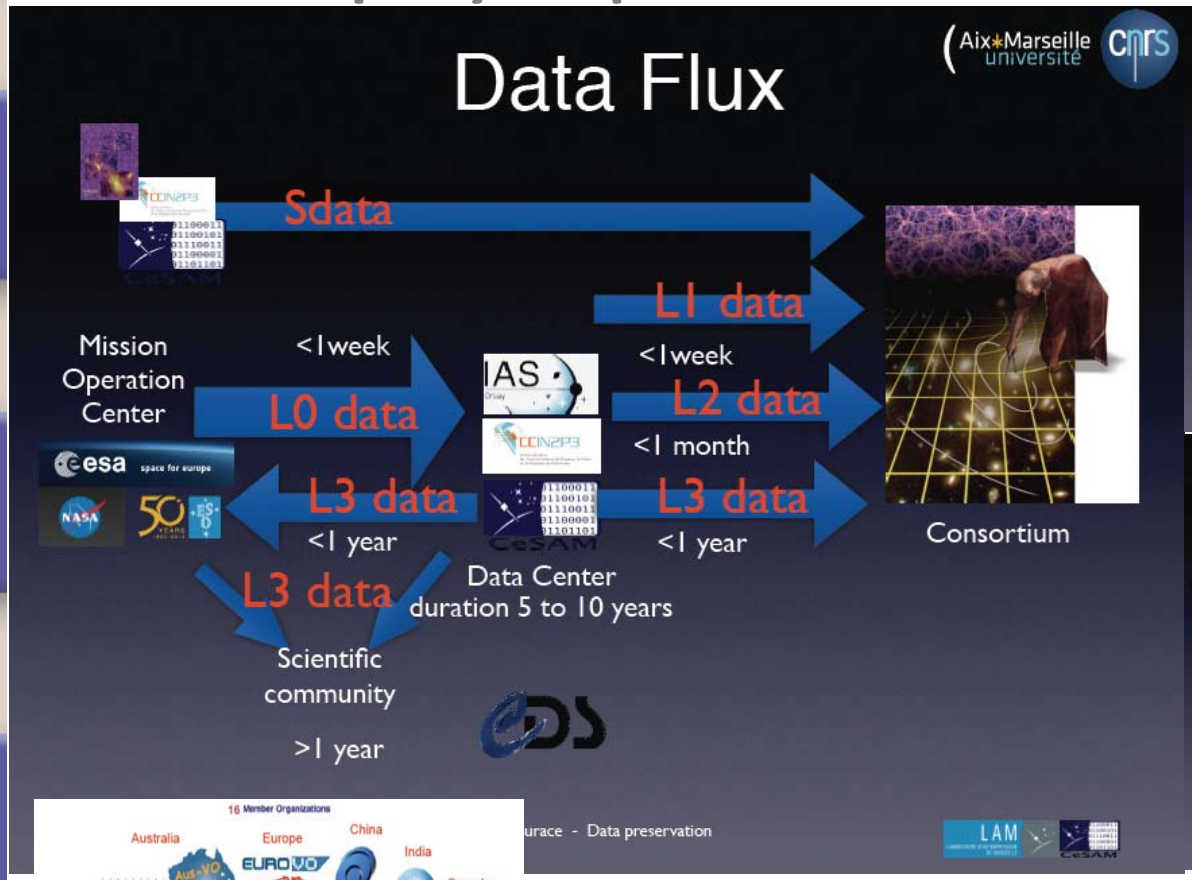
DPHEP Study Group for Data Preservation and Long Term Analysis in High Energy Physics

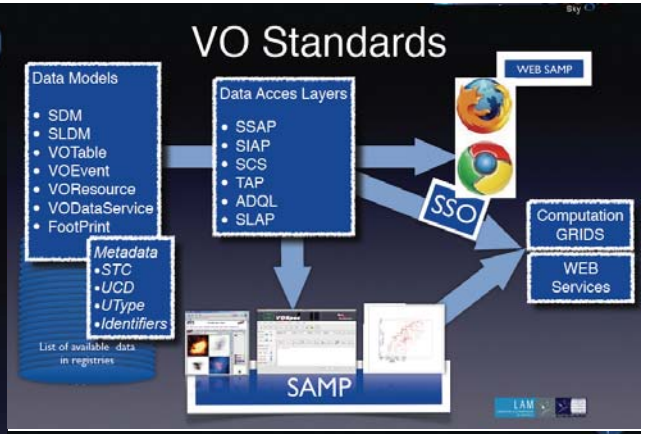
> **Organisation internationale**

MoU signé en juillet 2014

http://dphep.org

Astrophysique: Observatoires Virtuels

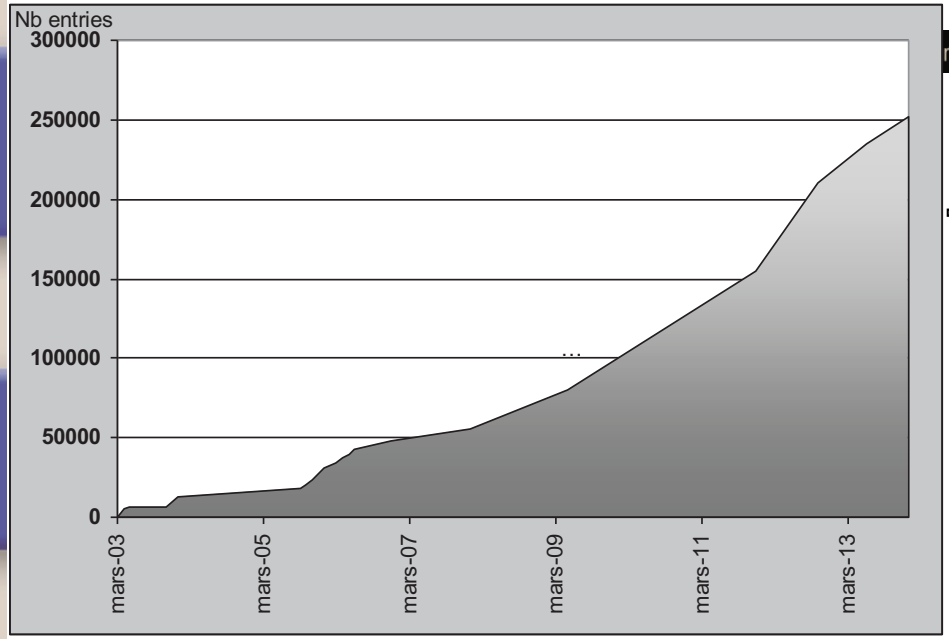


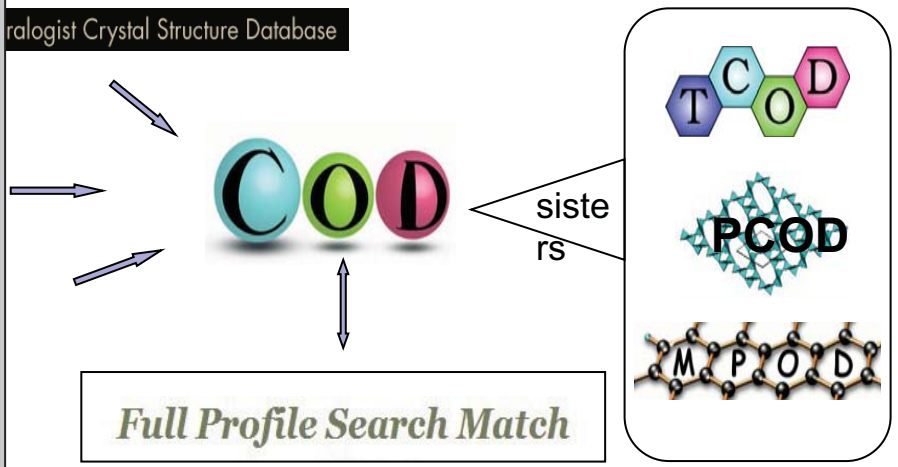


http://www.ivoa.org

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;

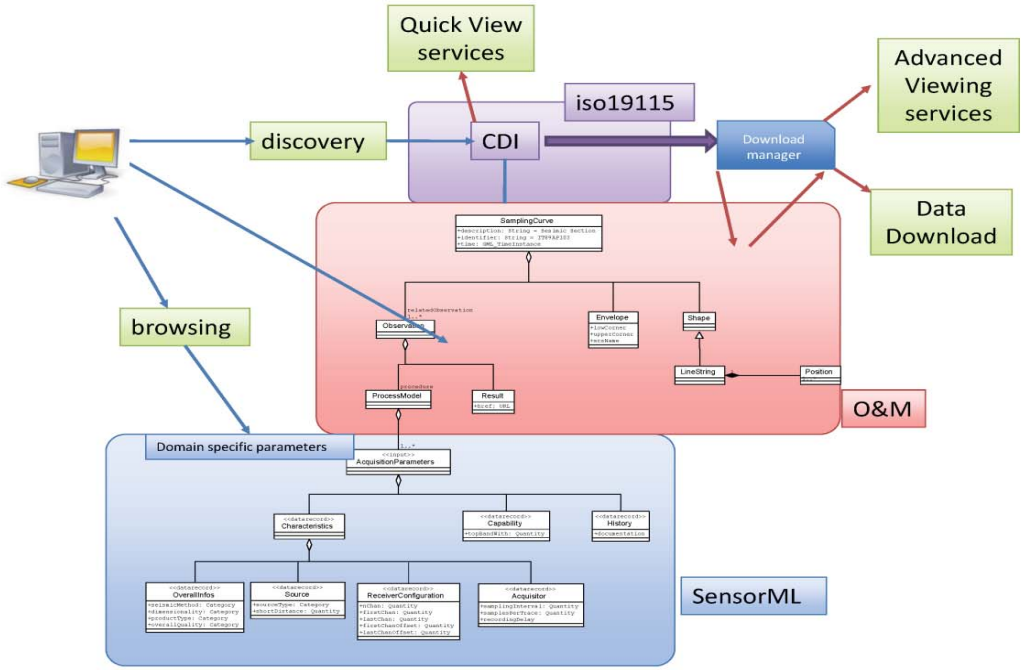
-few institutional data repositories yet have the expertise or resources to store the large quantities of data involved with the appropriate level of discoverability and linking to derived publications."

Seismic Data Preservation

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

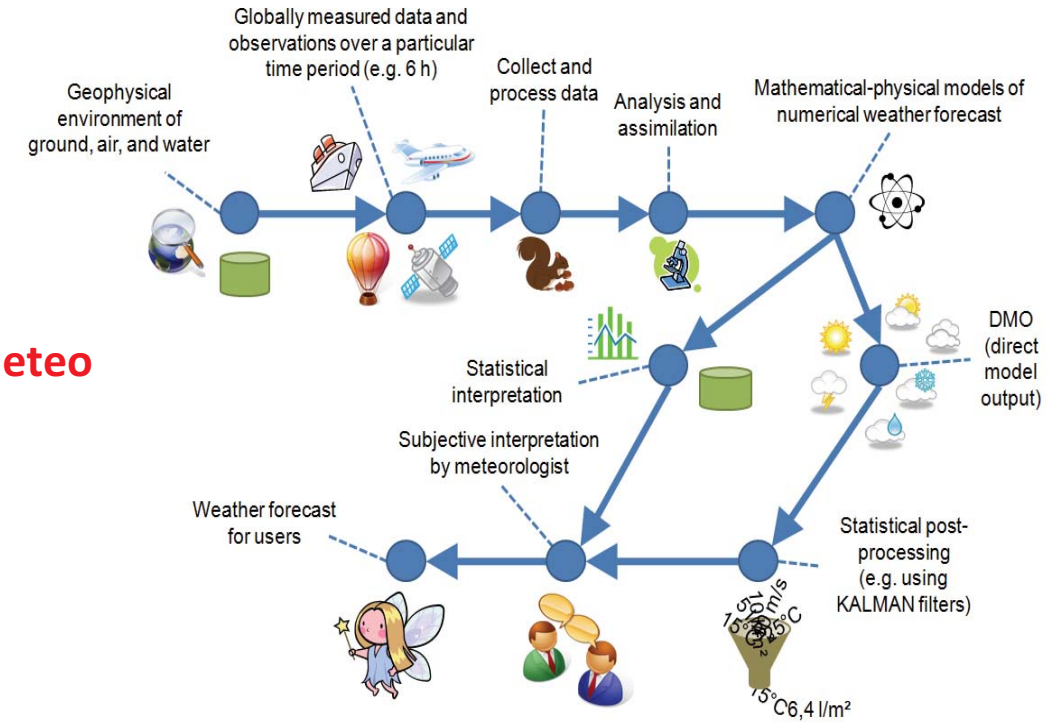
Conclusion

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.

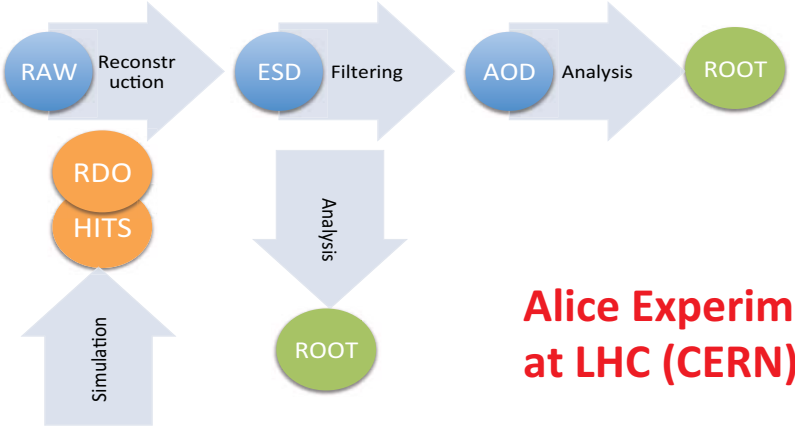


Formats, workflows et préservation

Meteo



=



Alice Experiment at LHC (CERN)

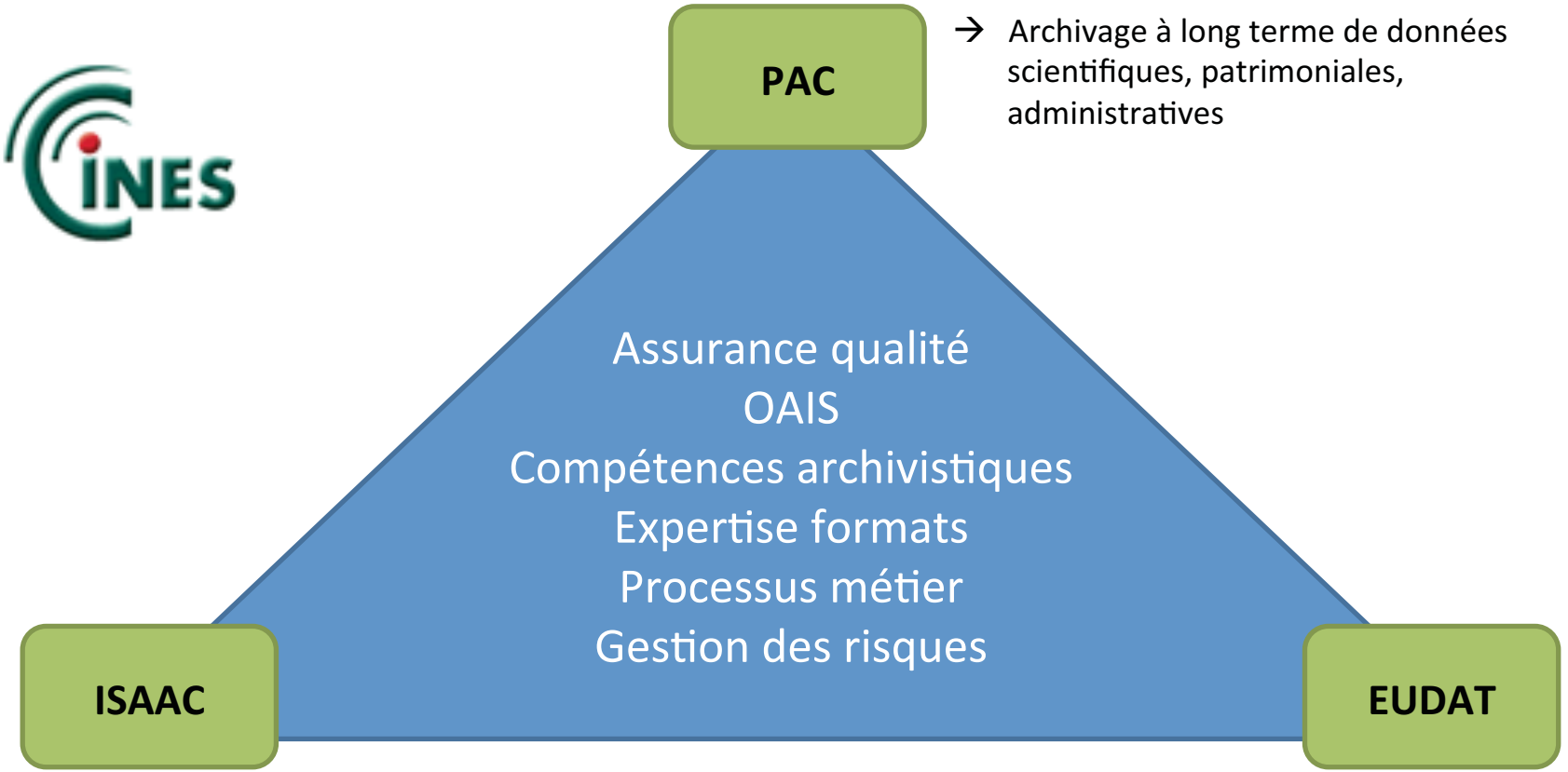
Formats de données: standards?

Similarité entre les disciplines

Approche théorique rigoureuse
Besoin et opportunité

Expertise au CINES

Les services d'archivage au CINES



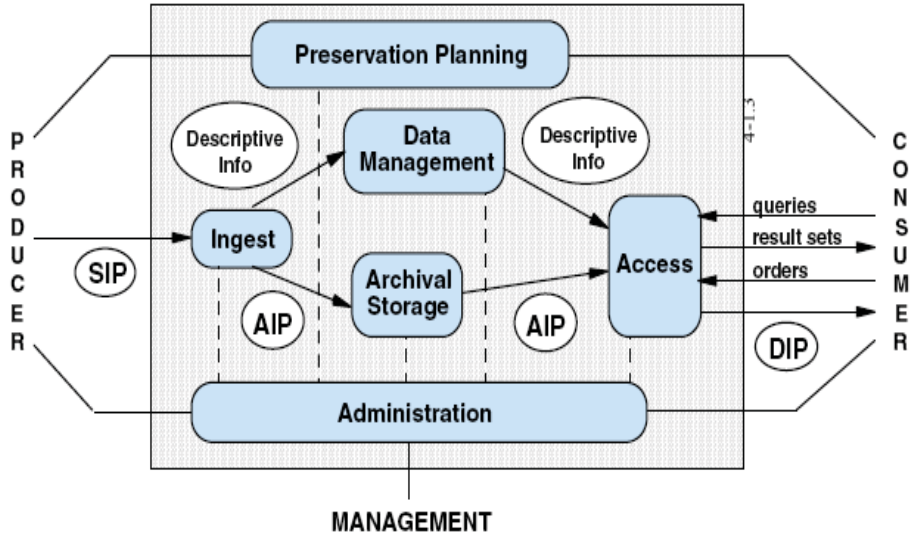
→ Archivage à long terme de données scientifiques, patrimoniales, administratives

→ Archivage intermédiaire de données scientifiques

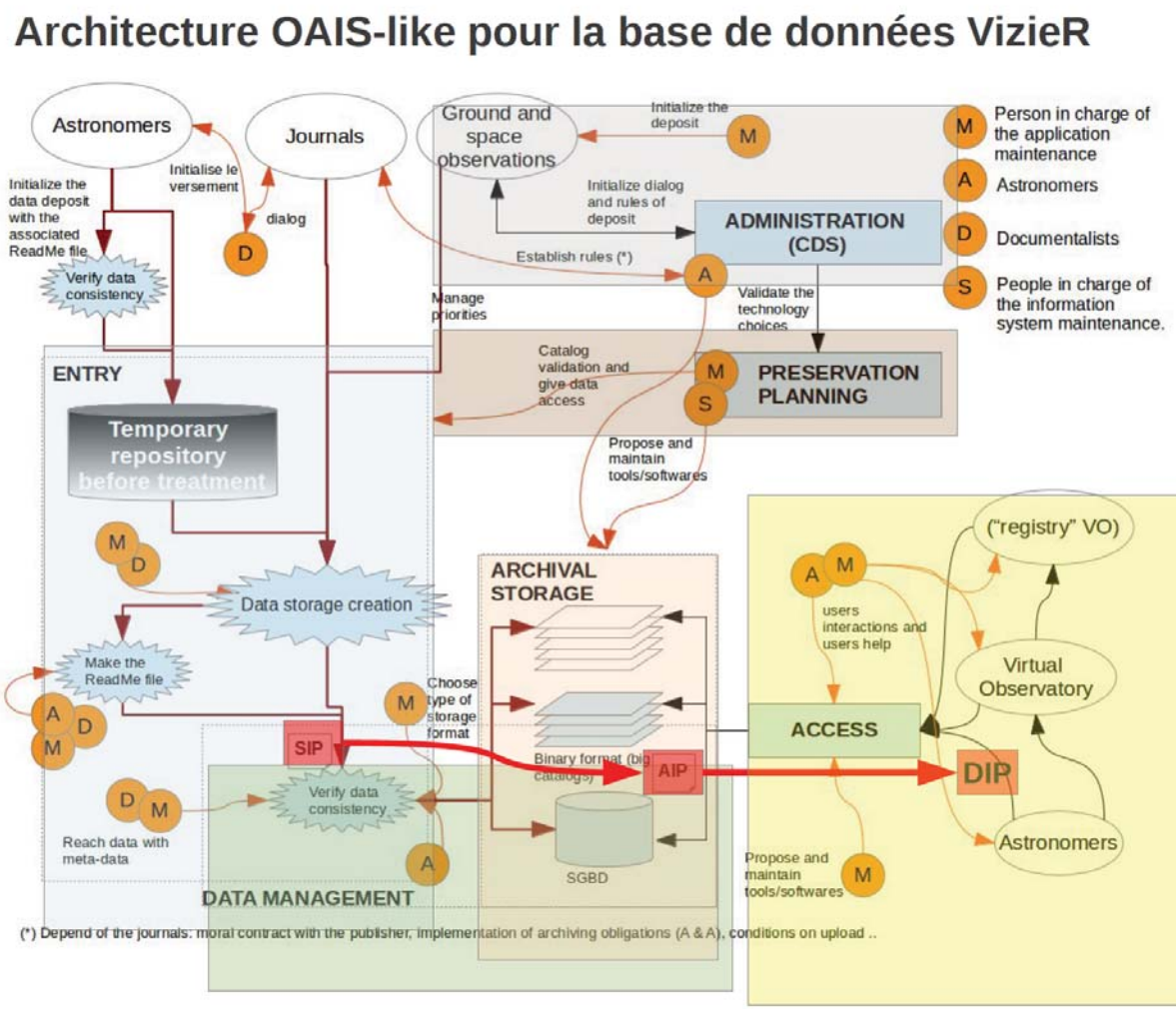
→ Archivage de données scientifiques pour des communautés européennes structurées

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



Centre de données de Strasbourg



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

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.

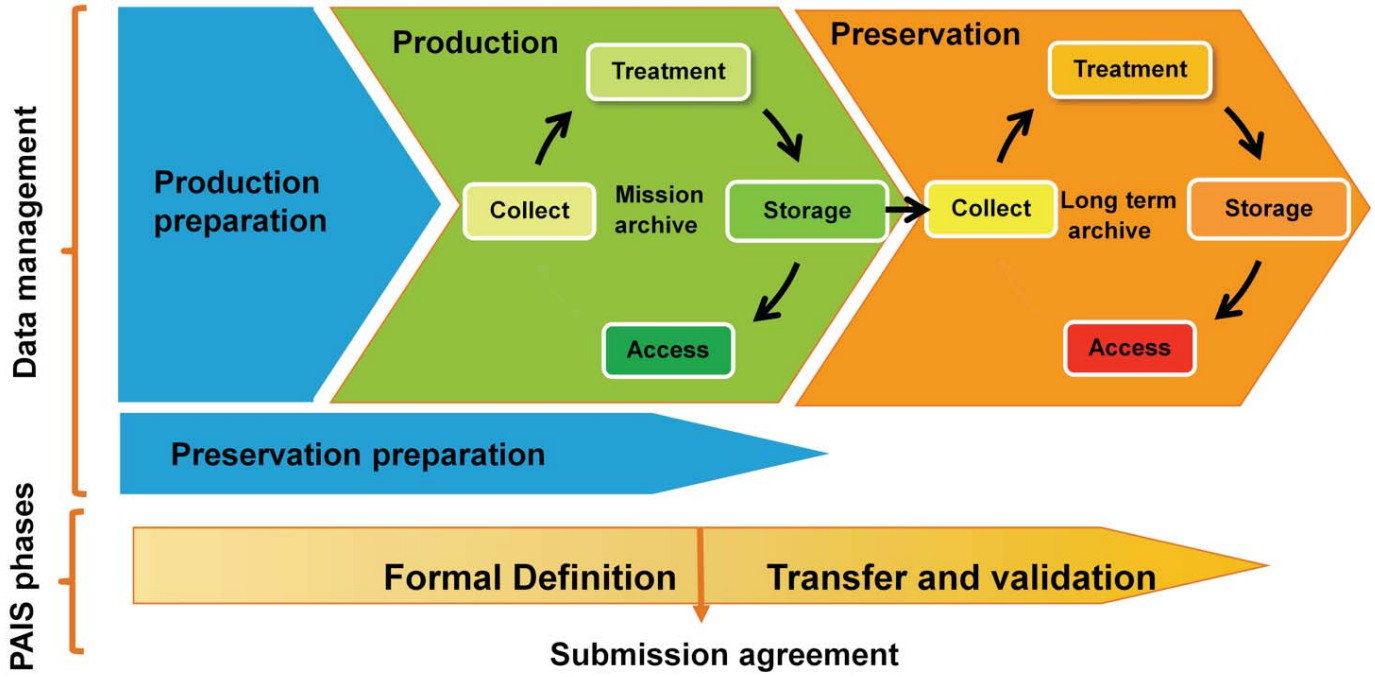
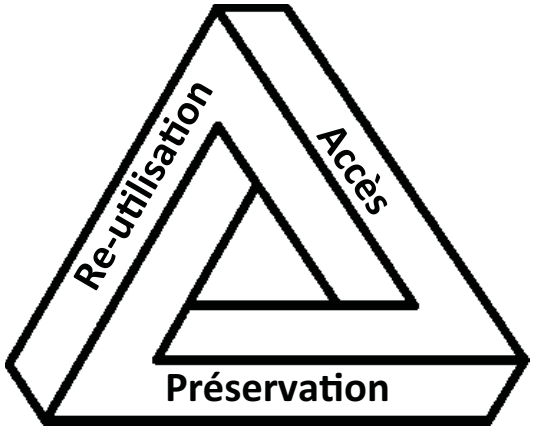


Figure 3: PAIS, preservation process and data lifecycle

Préservation, réutilisation, libre-accès



→ La préservation suppose la mise à disposition en accès libre

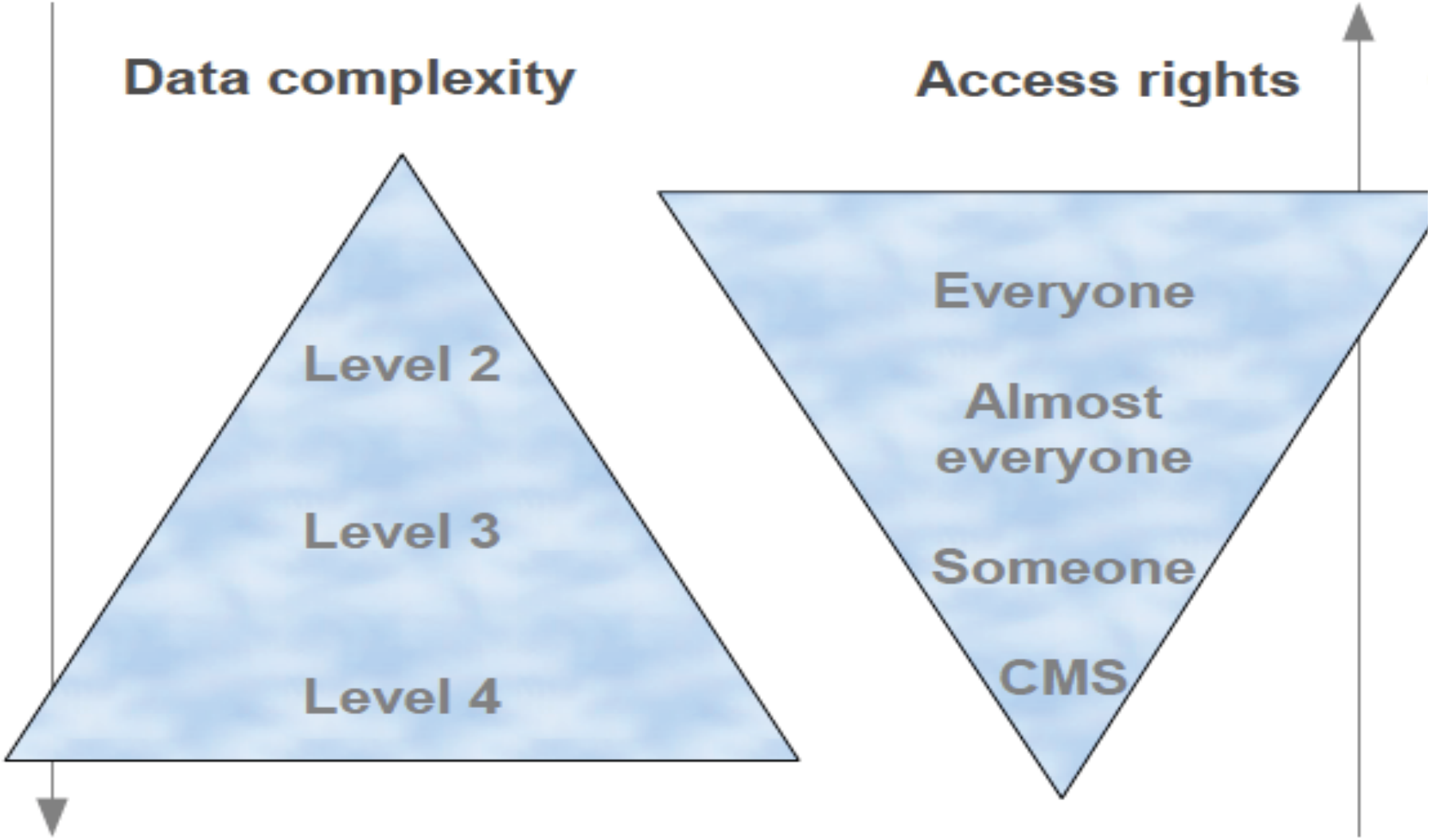
→ Maximiser le bénéfice

← Le libre-accès facilite la préservation à long terme

← Elargir la communauté, multiplier les connaissances

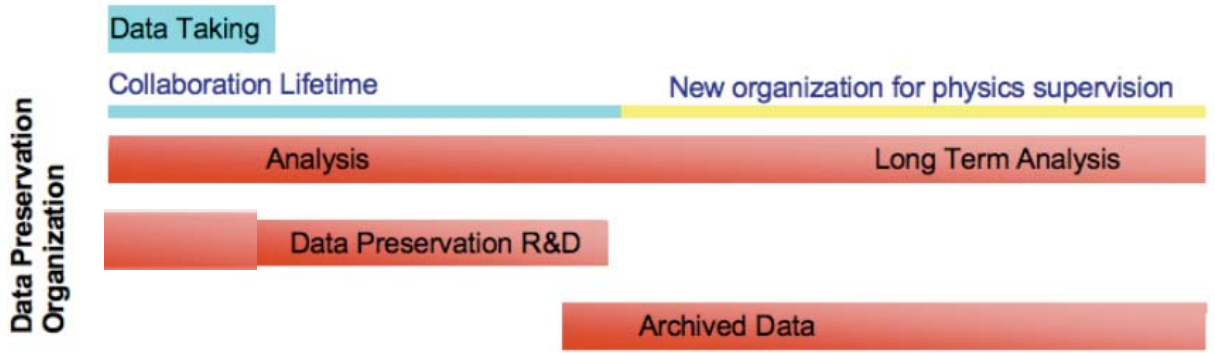
La propriété et les droits sur les données scientifiques à long térme?

Preservation complexity levels and access

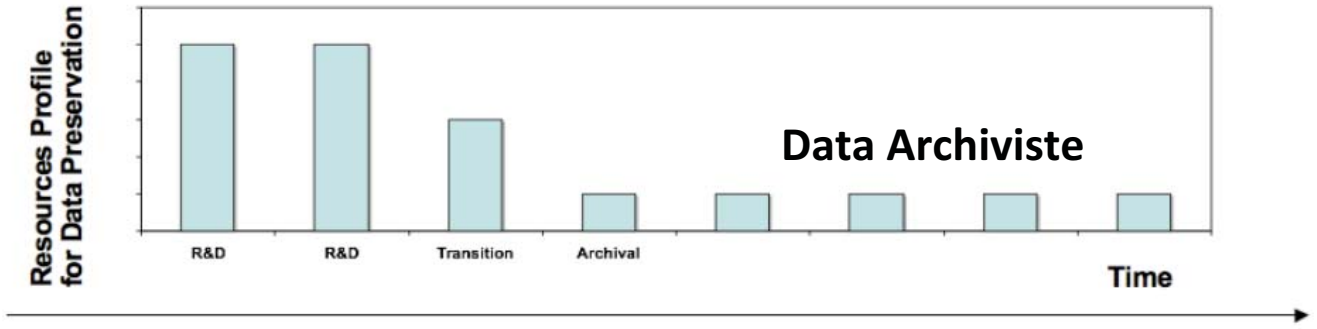


Organisation et ressources

ORGANISATION



RESSOURCES



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

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