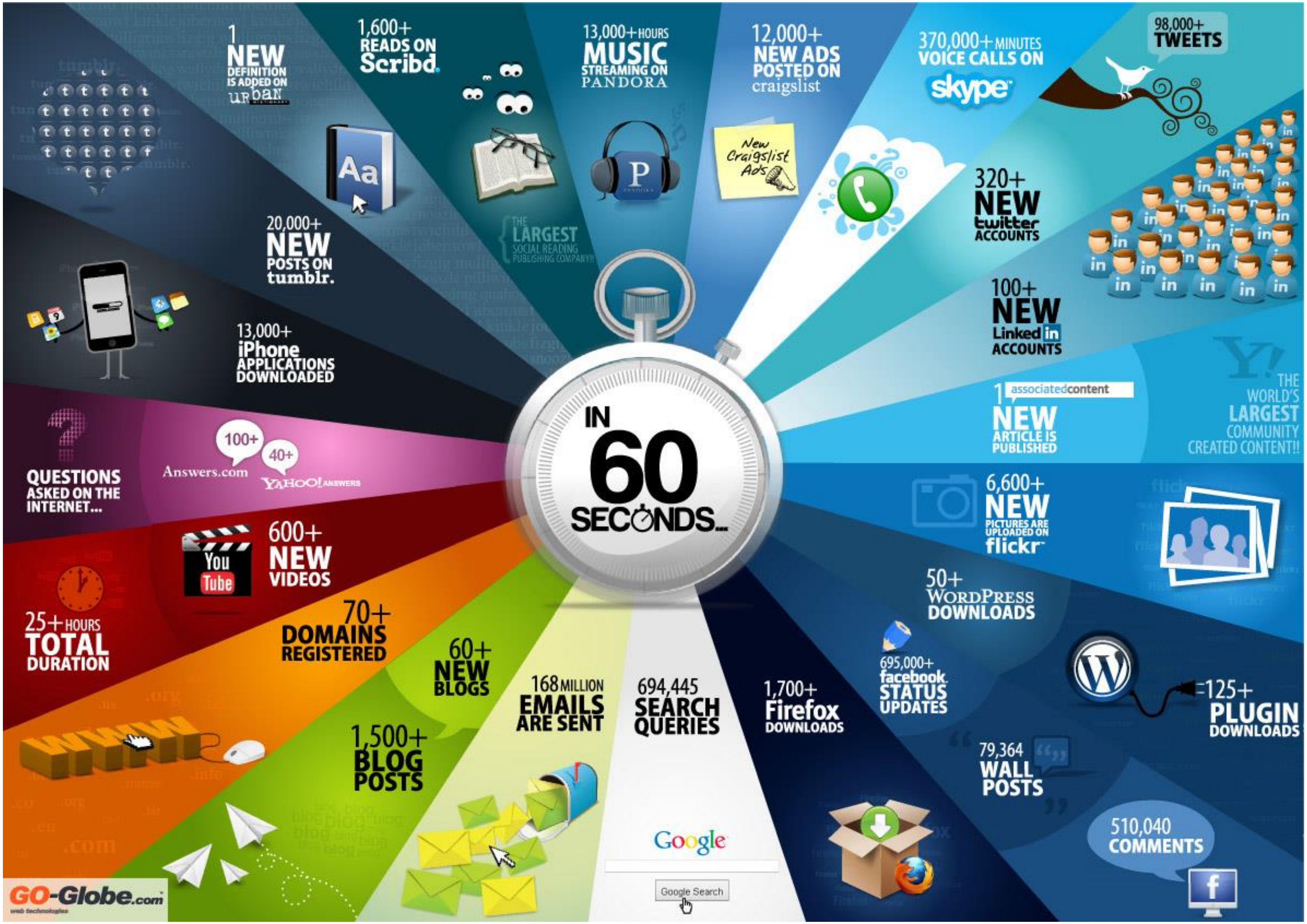
PREDON

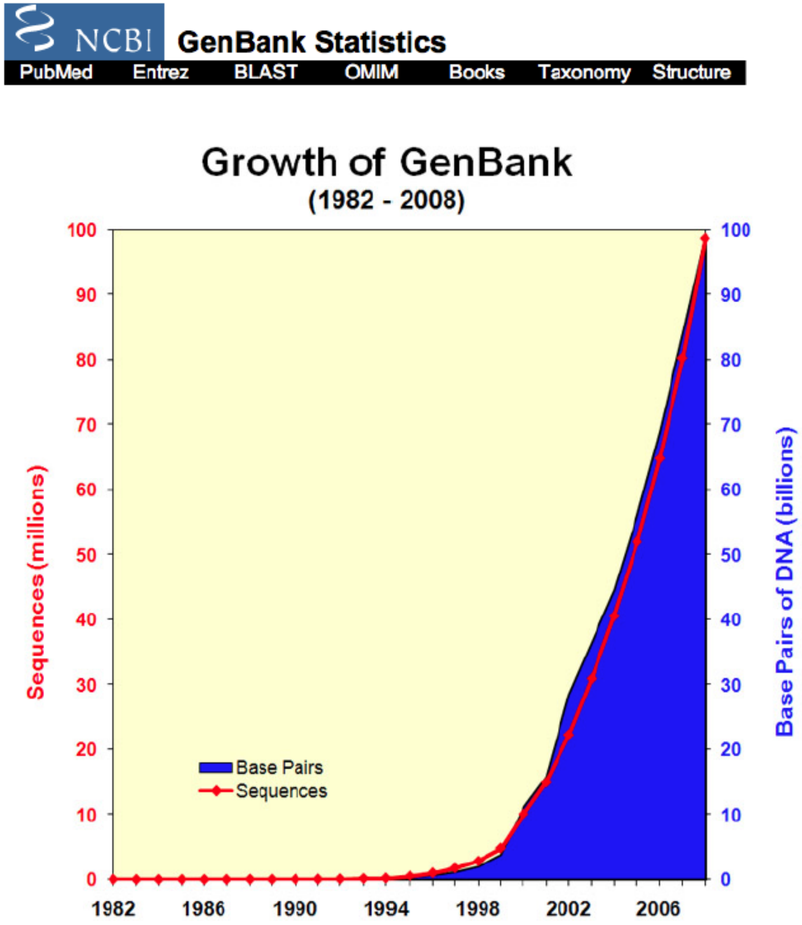
A project on Scientific Data Preservation in France
within MASTODONS multi-disciplinary program

Data Big Bang

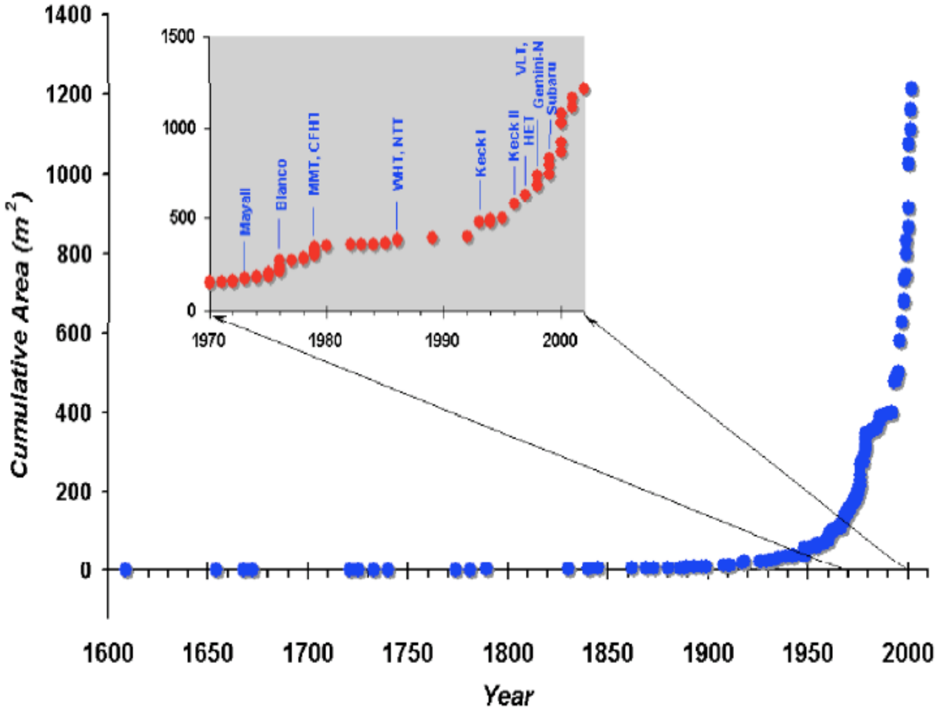

IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd.
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
New Craigslist Ads
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW LinkedIn ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WORDPRESS DOWNLOADS
125+ PLUGIN DOWNLOADS
695,000+ facebook STATUS UPDATES
79,364 WALL POSTS
510,040 COMMENTS
1,700+ Firefox DOWNLOADS
694,445 SEARCH QUERIES
Google
Google Search
168 MILLION EMAILS ARE SENT
1,500+ BLOG POSTS
60+ NEW BLOGS
70+ DOMAINS REGISTERED
25+ HOURS TOTAL DURATION
600+ NEW VIDEOS
You Tube
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
13,000+ iPhone APPLICATIONS DOWNLOADED
20,000+ NEW POSTS ON tumblr.
GO-Globe.com web technologies

Big Scientific Data

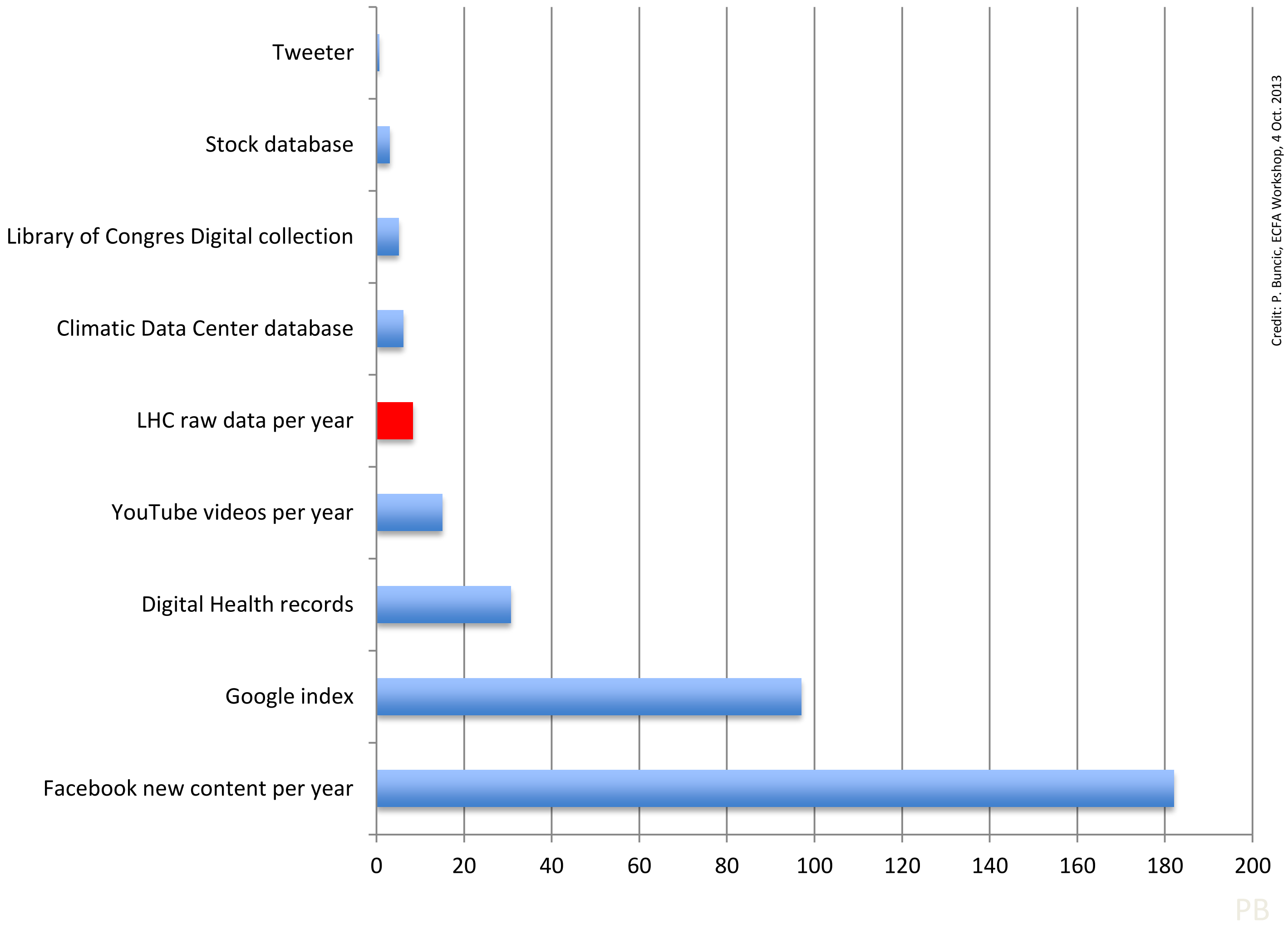
- Scientific research observes a dramatic increase in data and are questioning the long term future of this data



Telescope Collecting Area



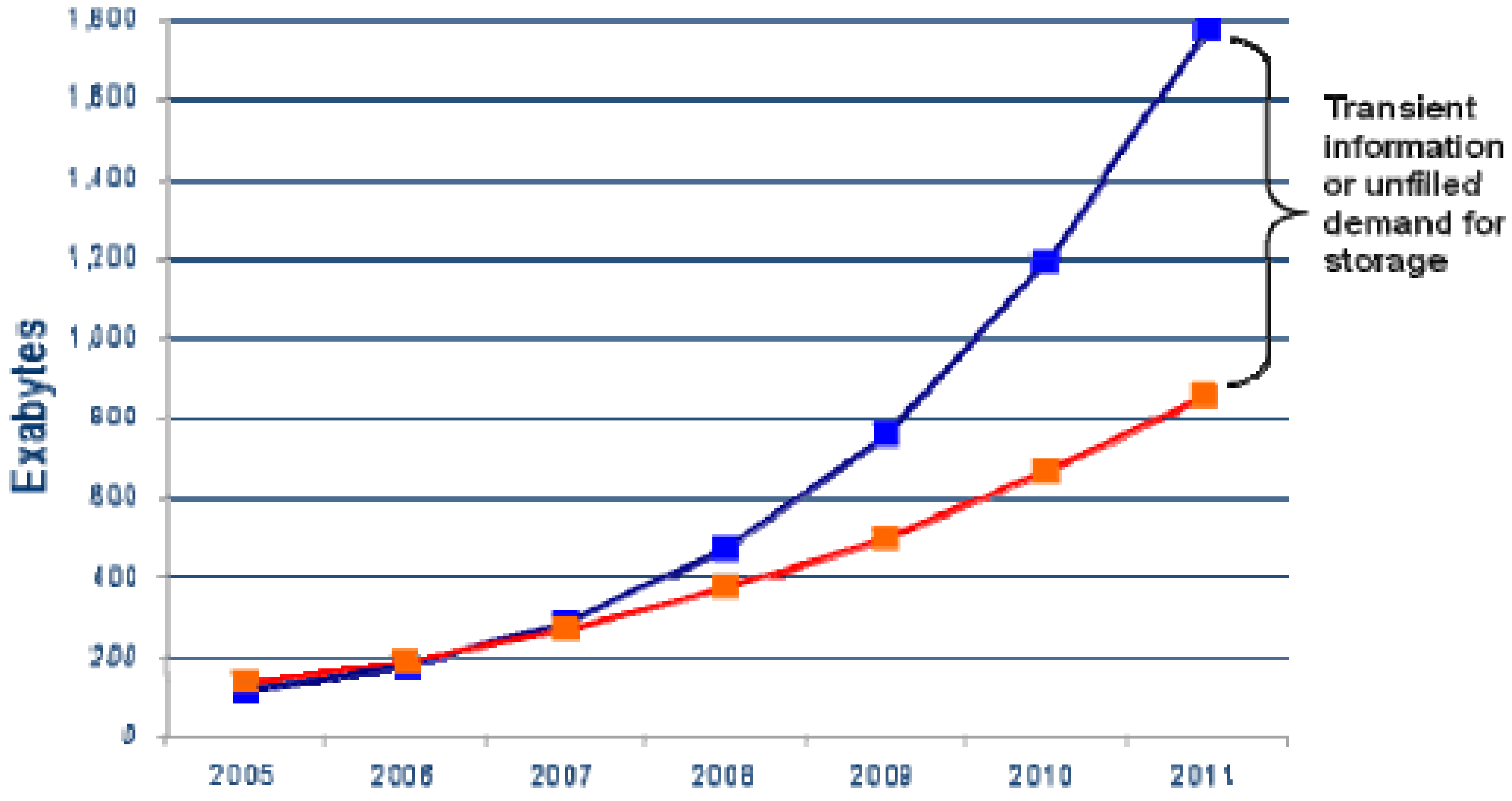
Big data: explosion des données digitales



Credit: P. Buncic, ECFA Workshop, 4 Oct. 2013

Les données digitales sont fragiles

- La capacité de stockage est physiquement dépassée depuis longtemps



Transient information or unfilled demand for storage

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

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

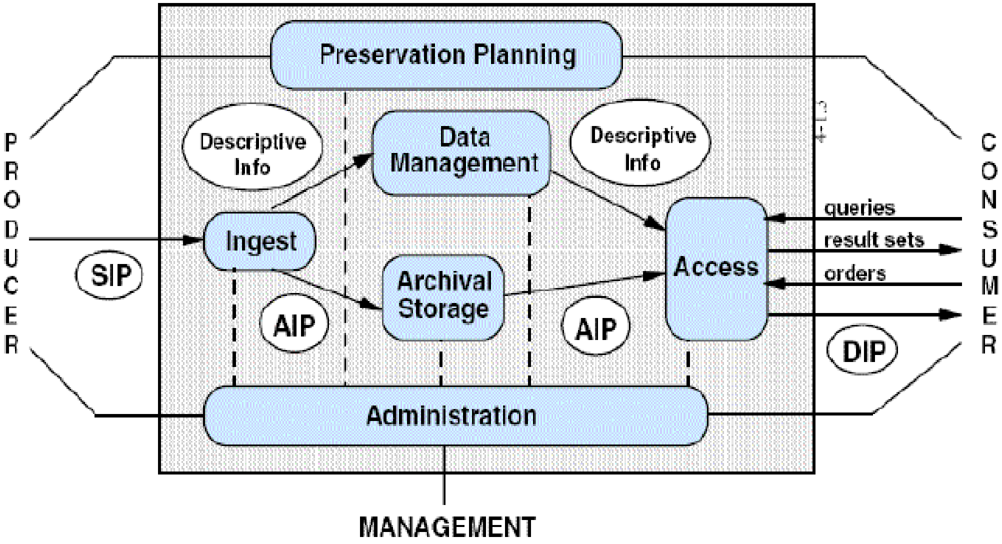


FIGURE 2.1: **The OAIS Reference Model**
http://public.ccsds.org/publications/archive/650x0b1.pdf, Page 4-1.
Source: Consultative Committee for Space Data Systems January 2002.

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

Est-ce que les données scientifiques sont spéciales?

- Riches en information car structurées suivant un plan de recherche et une démarche scientifiques
- De plus en plus diverses, la plus part des disciplines se sont mis a produire massivement des données
- Souvent produites avec des efforts financiers et humains significatifs (voir gigantesques)
  - Plus ca 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
- Il est évident qu'on doit réfléchir (à deux fois) sur le sort de ces données
  - PRESERVATION!

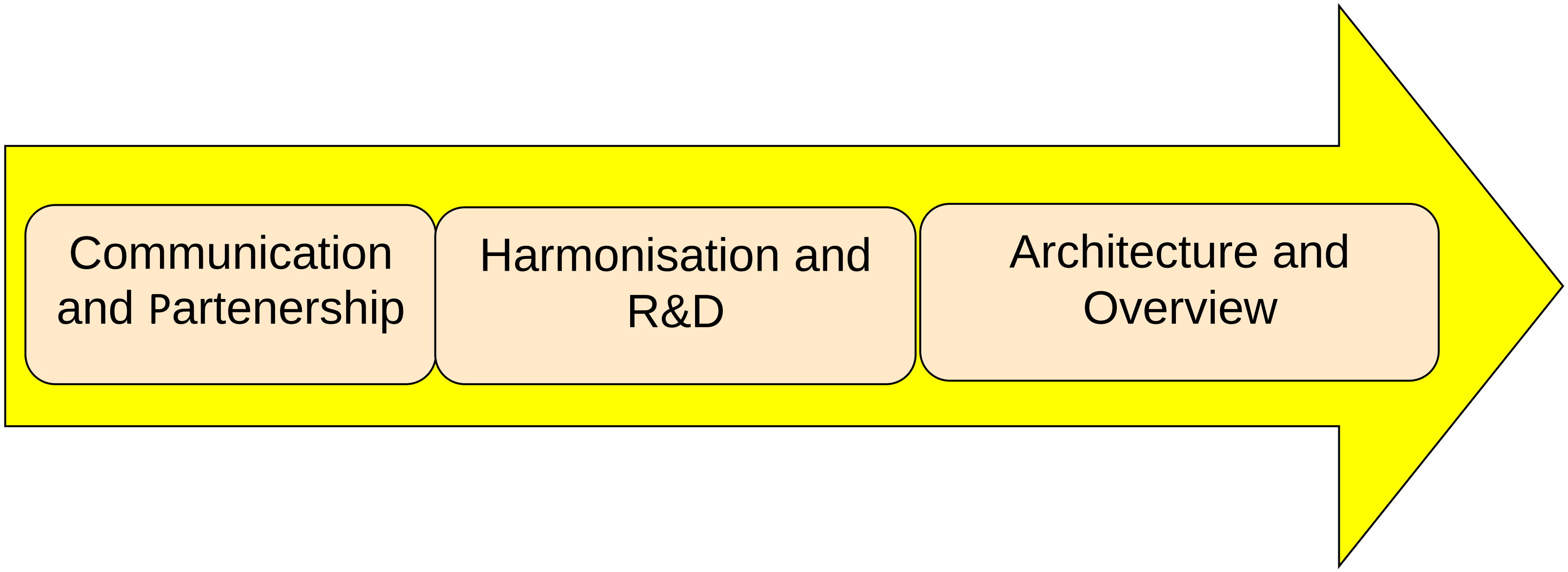
MASTODONS

- Multi-disciplinary Department of CNRS launched a call in April 2012:
  - Data exceeds storage
  - More science in Data

- Possible Directions:
  - 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)**

- MASTODONS is likely to be evolved in a national program around big data

PREDON: Plans



Communication and Partenership

Harmonisation and R&D

Architecture and Overview

- Short term (2013/2014): **Communication and partenership**
  - Enlarge the community
- Medium term (2014/2015) : **Harmonisation and R&D**
  - Communication: exchanges and workshops
  - Demonstrator acces and préservation
- Long term (2016) **Architecture and overview**
  - "Observatoire National des Données Scientifiques"

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.

Consortium PREDON

- Formation d'un consortium avec des compétences complémentaires

- Physique des particules, astroparticules et théorie
  - CPPM IN2P3 PP
  - LAPP (IN2P3) astro-particules
  - LPSC (IN2P3) physique theorique

PREDON Avril 2012

- Astrophysique
  - APC/FACe (IN2P3) astroparticules, astrophysique
  - OAMP/LAM (INSU) astrophysique

- Recherche informatique: exploitation des grandes masses de données complèxes
  - LIRMM (INS2I) Univ. Montpellier
  - Univ. Paris 5
  - Univ. Paris 13
  - Espace DEV (UM2 +IRD UAG ULR)

- Grands centres de calcul
  - CC-IN2P3 Centre de calcul IN2P3
  - CINES Centre Informatique National de l'Enseignement Supérieur
- Contacts en cours: CNES, ExaBuilder

PREDON Décembre 2012

PREDON Consortium

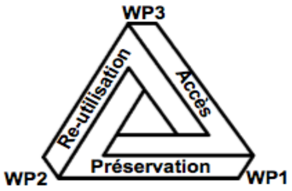
	Volume données	Complexité	Diversification des sources	Structuration au niveau international	Algorithmes et methodologies pour la preservation
IN2P3 HEP	+++	+++	+	++	+
INSU, IRD Astrophysics Earth Sciences	++	++	++	+++	++
CINES INS2I IT, Algorithms, workflows	+	++	+++	+	+++

Nouveau contacts:
Daniele Boucon, expert en preservation de données CNES
Daniel Chateigner, CRISMAT/ENSICAEN, données cristallographie
Catherine Boisson de l'Observatoire de Meudon / LUTH/INSU CTA

PREDON: Objectives

- **Identification** of the scientific and technical **requirements** for an unified approach of data preservation within an multi-disciplinary context.
  - *IOS:* installation of an unified platform to store at long term scientific data in a multidisciplinary context. A demonstrator is proposed within this project.
- Reinforcement of the **coherence and standardisation** of data collection, storage, analysis and access in several scientific domains with complementary needs, leading to a robust and friendly environment for long-term data preservation.
  - *IOS*: installation of a multi-disciplinary mechanism for data preservation standards
- Installation of a **scientific data tracking and supervision system**, such that the information produced during the scientific experiments is followed and centrally tagged at all stages: production, exploitation, archival.
  - *IOS:* definition of a national organisation relative to the scientific data preservation, aimed at supporting main experimental scientific branches producing scientific data towards a traceable long term data preservation
- Reinforcement of **the international cooperation** on this issue in a context of a vast effort to treat large amounts of data sets.
  - *IOS:* permanent links of the consortium with the corresponding international organisations (for instance DPHEP) and the participation of the consortium to the relevant European programs in the field (for instance programs included in the Horizon 2020 agenda).

PREDON as a project

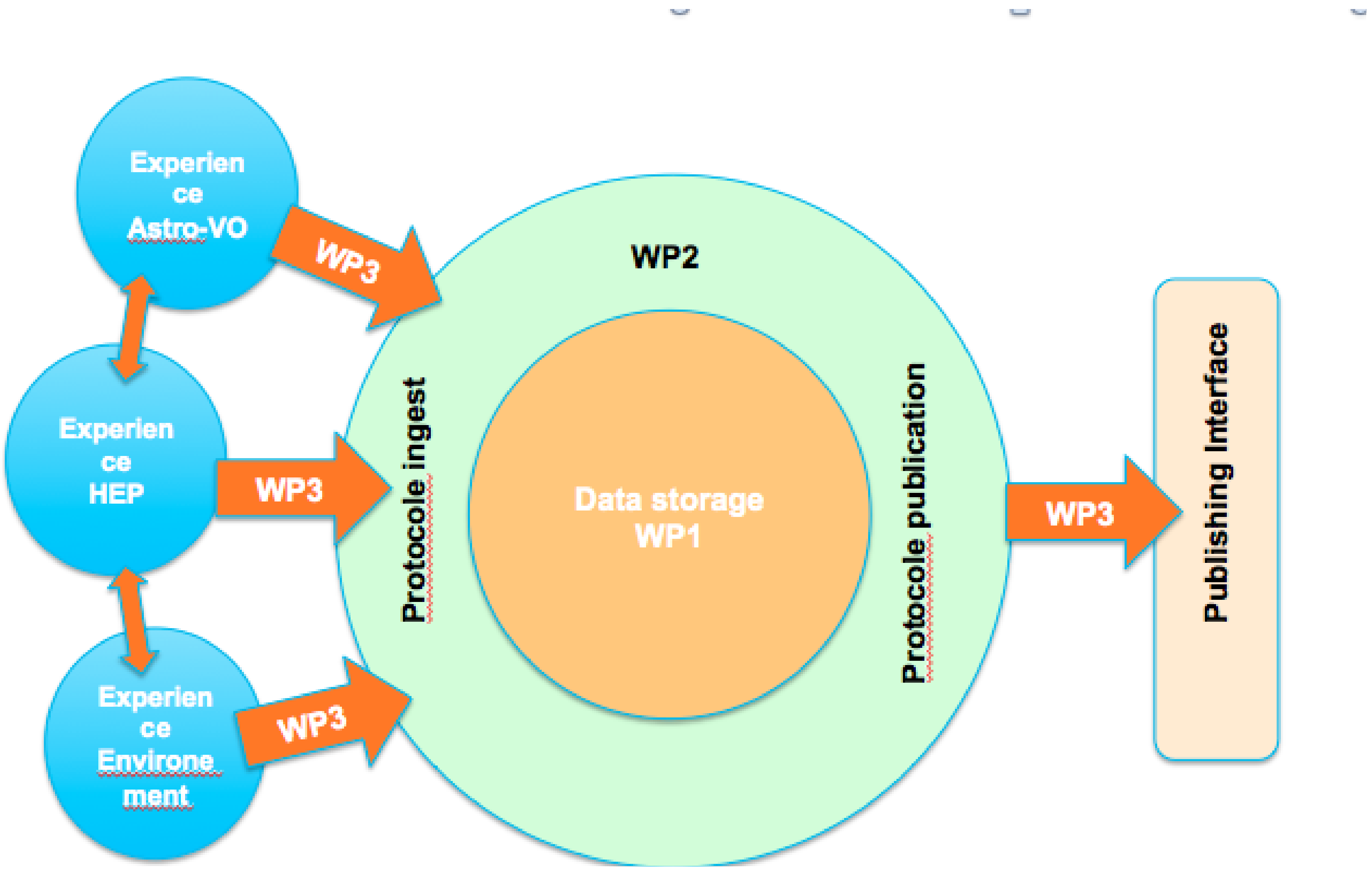


Working Package	Objectives	Participants (*coordinator)
WP1 Technologies and Methodologies	Explore methodologies and technologies suitable for a coherent and robust scientific data preservation in a multi-disciplinary context and on a multi-platform computing centre	CINES* APC
WP2 Algorithms and Workflows	Investigate generic and mathematically robust workflows and algorithms for data mining suited for data and workflow preservation; data- and process-based workflows and mining techniques to be used in a multi-disciplinary environment towards long term data preservation	LAM LIRMM LIPADE* LIPN
WP3 Data formats and interfaces	A parallel approach for data collection, storage, processing, analysis and preservation with the aim to achieve common standards for scientific data treatment	APC CPPM LAM* LPSC
WP4 General coordination	Program coordination, dissemination and international cooperation	CPPM*

Objectifs scientifiques 2013 (dec. 2012)

Objectif	Moyens	Resp.	Delivrables
Renforcement de la communication entre les partenaires, extension du consortium au niveau FR [connexions MASTODONS]	1 atelier (généraliste) en France Mise en place des outils collaboratif	WP0	Actes des rencontres, compte- rendus des conclusions Site web, forge, espace développement
Renforcement des connection internationales	Participation aux réunion de travail RDA et EUDAT (etc.)	WP0	Proposition de financement communes, participation a des appel d'offre et constitution de consortia
Exploration d'un demonstrateur de stockage intégré des données scientifiques	Serveur de données dédié, réunions de travail ciblées avec des experts	WP1	Note technique sur la mise en place du serveur, les methodes et les résultats;
Méthodes et algorithmes d'indexation et préservation des données scientifiques	Réunions de travail entre experts CINES/CC-IN2P3. Connections producteur de données(DPHEP/LHC, VO, EO)	WP2	Livre blanc de recommendations et procédures; procedure demonstrative de stockage des données complexes suivant le protocole.
Standardisation des formats et des modèles de description bi- et multi-disciplinaires: données et accès	Réunions a distance, stages de travail	WP3	Publication d'un prototype de format unique dans PHE (niveau à déterminer) suivant la méthodologie utilisée dans l'astrophysique. Projet Outreach. Pistes pour une logique intégré suivant le cadre du projet ISAAC.

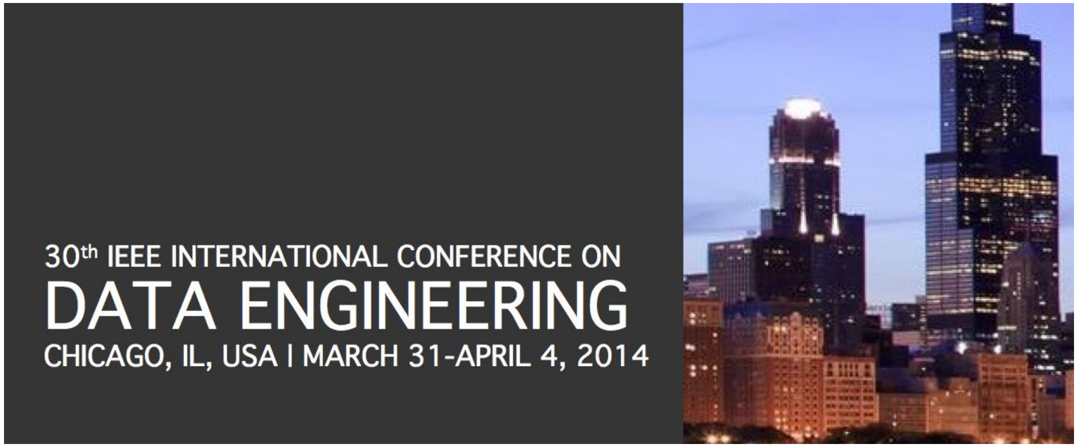
Demonstrator



Workshop on Data Preservation at ICDE 2014

LOPS@ICDE

WORKSHOP ON LONG TERM PRESERVATION FOR BIG SCIENTIFIC DATA



30th IEEE INTERNATIONAL CONFERENCE ON

DATA ENGINEERING

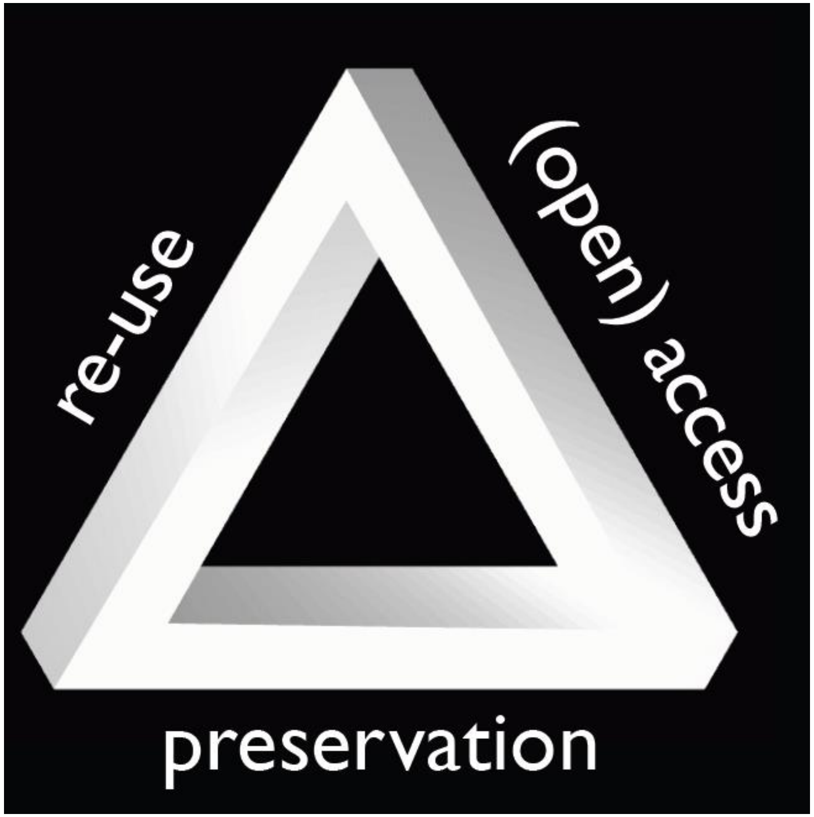
CHICAGO, IL, USA I MARCH 31-APRIL 4, 2014

- 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.
- Paper submission deadline November 10, 2013

A word on access and data preservation

Example: NSF Policy
Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples,
physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing.

Proposals […] must include a supplementary […] "Data Management Plan" (DMP) […] describ[ing] how the proposal will conform to NSF policy on the
dissemination and sharing of research results.
http://www.nsf.gov/bfa/dias/policy/dmp.jsp



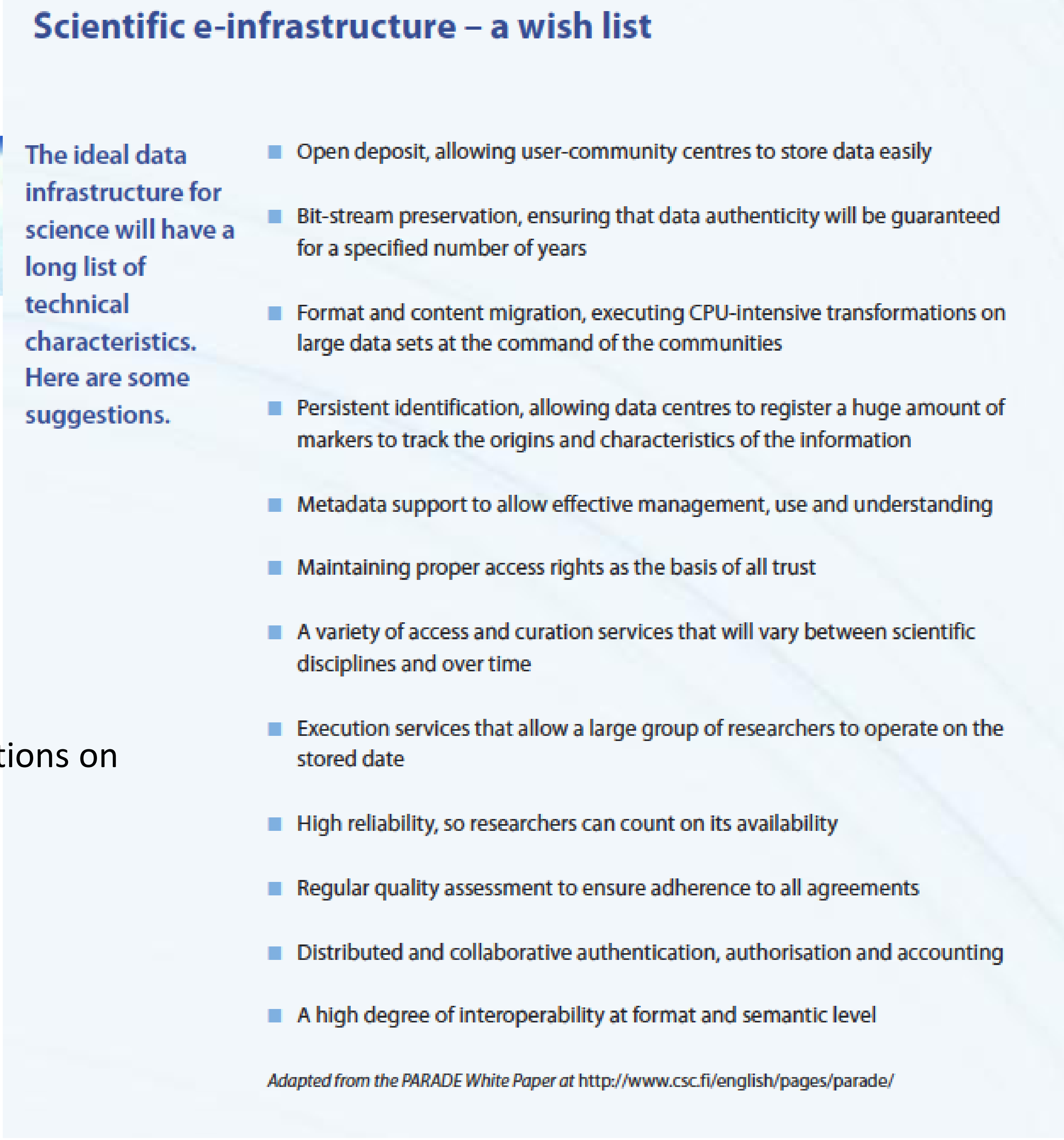
Very similar policies in other funding agencies (and growing interest for these aspects in the context of “big data” strategies)

Riding the wave

How Europe can gain from the rising tide of scientific data

Final report of the High Level Expert Group on Scientific Data
A submission to the European Commission

October 2010



Scientific e-infrastructure – a wish list

The ideal data infrastructure for science will have a long list of technical characteristics. Here are some suggestions.

- Open deposit, allowing user-community centres to store data easily
- Bit-stream preservation, ensuring that data authenticity will be guaranteed for a specified number of years
- Format and content migration, executing CPU-intensive transformations on large data sets at the command of the communities
- Persistent identification, allowing data centres to register a huge amount of markers to track the origins and characteristics of the information
- Metadata support to allow effective management, use and understanding
- Maintaining proper access rights as the basis of all trust
- A variety of access and curation services that will vary between scientific disciplines and over time
- Execution services that allow a large group of researchers to operate on the stored date
- High reliability, so researchers can count on its availability
- Regular quality assessment to ensure adherence to all agreements
- Distributed and collaborative authentication, authorisation and accounting
- A high degree of interoperability at format and semantic level

Adapted from the PARADE White Paper at http://www.csc.fi/english/pages/parade/

A myriad of projects/coalitions on
“data infrastructures”
either funded or
in preparation for FP8

-APA, EUDAT, DPM, RDA…

RDA Preservation WG

- The RDA – strongly supported by EU, NSF, AU – seen as an element of implementing HLEG 2030 vision
- A Interest Group on DP was approved in May
  - Chair: David Giaretta (APA, SCIDIP-ES, author of “Advanced DP”, ex-DCC, ex-STFC)
  - Co-chair, rapporteur: Jamie Shiers (PM DPHEP)
- The intent is to show progress by each RDA plenary (March, September) and co-ordinate international activities, identify candidate services for standardization, lobby for funding…

RDA
RESEARCH DATA ALLIANCE
Research Data Sharing
without barriers

RDA IG – Work steps (until DUB)

- **Regular virtual meetings**
  - **Contribute concepts:**
    - Use cases
    - Potential services + Relevant abstract interfaces
  - **Identify:**
    - where we can bring existing capabilities together – as proof of concept
    - “gaps” in shared preservation e-infrastructure *(to be filled via projects?)*
    - **how the work of other IGs and WGs can fit in**
    - **potential WGs arising from this IG**
  - **(Eventual) outcomes:**
    - **Preservation tool-kit, “Services”, e.g. media migration**

Buts du workshop PREDONx 2013

- Tour des projets au sein de PREDON
- Elargir le champ de communication sur le sujet DP aux autres projets Mastodons
- Nouvelles approches: documentation, juridique, économique
- Connexion aux projets similaires en France
- Document PREDON2013: papier blanc avec les conclusions générales du workshop

Backup

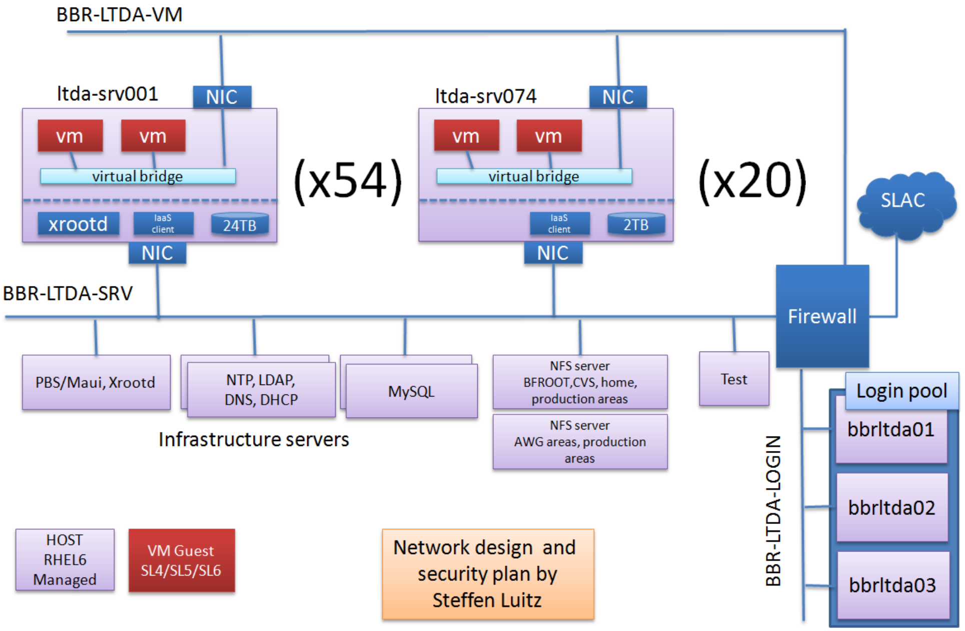
PREDON: Next Steps

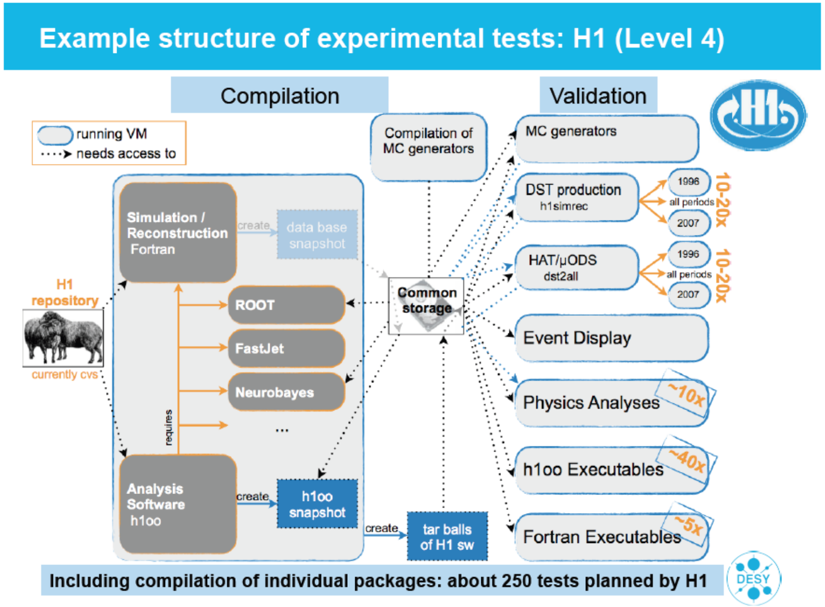
- More aspects
  - Scientific and technical information (libraries &co.)
  - Legal aspects
  - Economical models
- White paper end 2013/2014 to national funding agencies
  - Vol 1: Facts finding
  - Vol 2: Projects
  - Vol 3: Organization

Exemples projets PHE

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

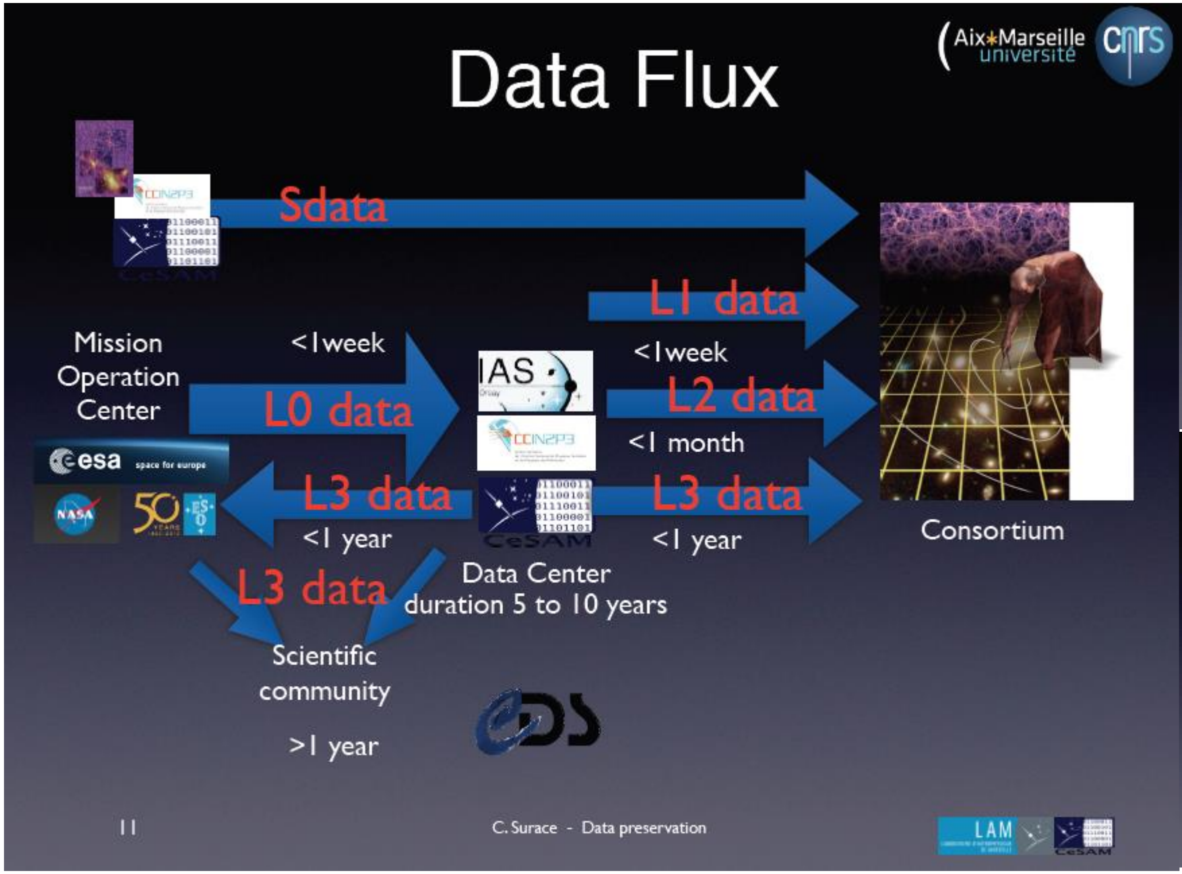
Système de préservation et migration
Virtualisation, validation intensive
(DESY, Hambourg, Allemagne)





DPHEP « Project Manager » nommé au CERN en Octobre 2012

Exemple projet astrophysique: Virtual Observatories



Data Flux

Sdata

L1 data

Mission Operation Center

<1week

L0 data

<1week

L2 data

<1 month

L3 data

<1 year

L3 data

<1 year

L3 data

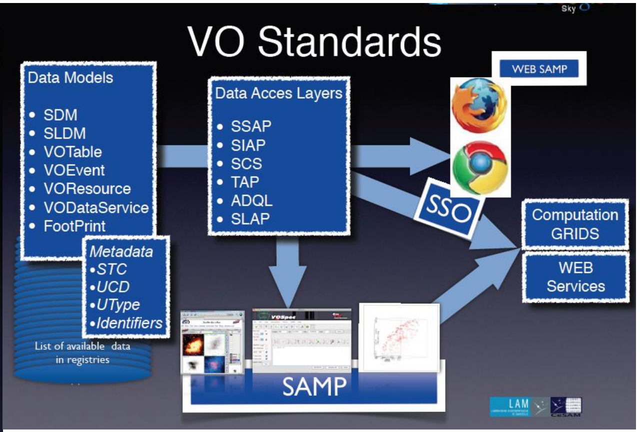
Scientific community

>1 year

Data Center duration 5 to 10 years

Consortium

11 C. Surace - Data preservation



VO Standards

Data Models

- SDM
- SLDM
- VOTable
- VOEvent
- VOResource
- VODataService
- FootPrint

Metadata

- *STC*
- *UCD*
- *UType*
- *Identifiers*

List of available data in registries

Data Acces Layers

- SSAP
- SIAP
- SCS
- TAP
- ADQL
- SLAP

WEB SAMP

SSO

Computation GRIDS

WEB Services

SAMP

VO - Tools

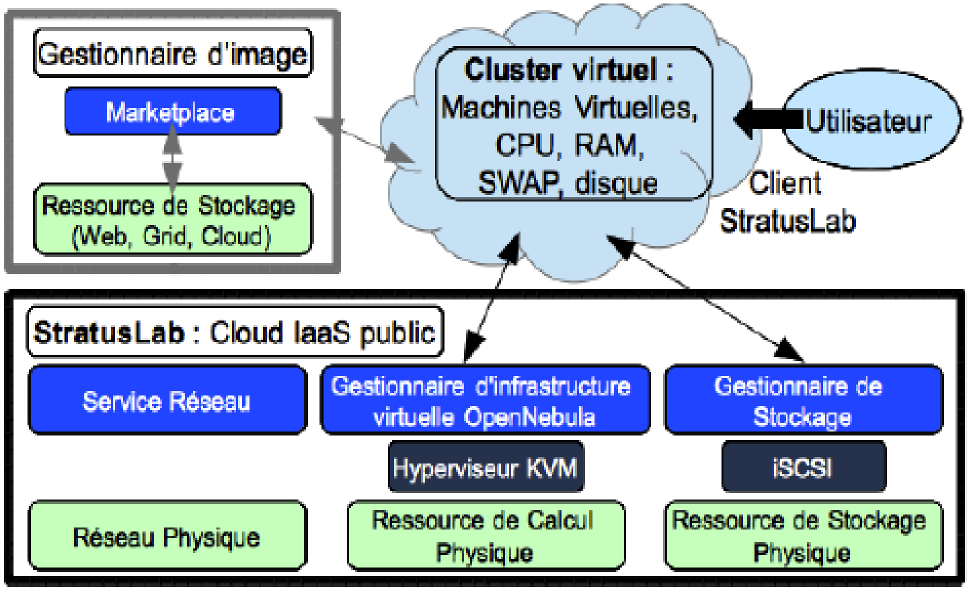
15 C. Surace - Data preservation

http://www.ivoa.org

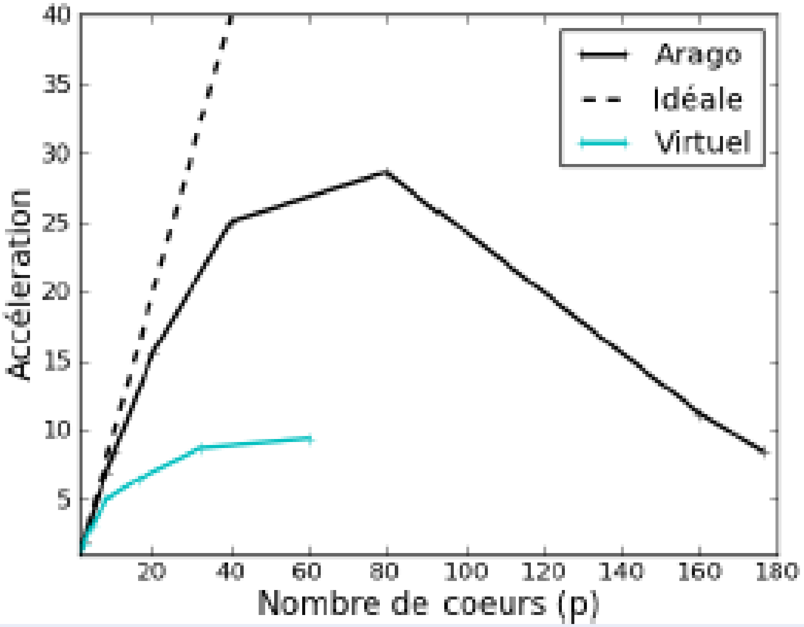
Exemple projet: Data processing & storage in the cloud

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
- questions: accessibility, data security, short-term and long-term cost


Gestionnaire d'image
Marketplace
Ressource de Stockage (Web, Grid, Cloud)
Cluster virtuel : Machines Virtuelles, CPU, RAM, SWAP, disque
Utilisateur
Client StratusLab
StratusLab : Cloud IaaS public
Service Réseau
Gestionnaire d'infrastructure virtuelle OpenNebula
Gestionnaire de Stockage
Hyperviseur KVM
iSCSI
Réseau Physique
Ressource de Calcul Physique
Ressource de Stockage Physique

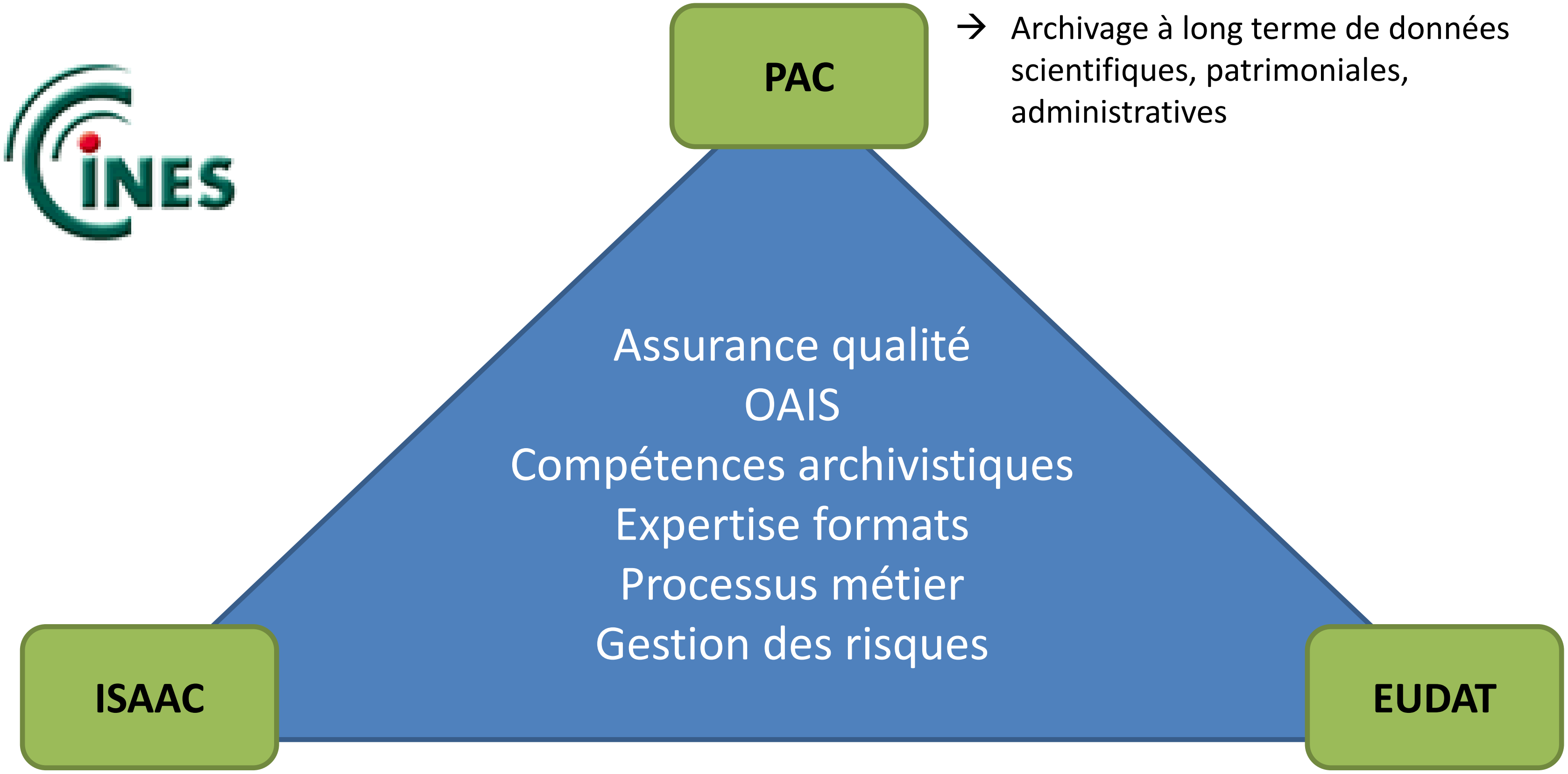
Schematic description of the cloud StratusLab, which is a European public cloud project IaaS which started in 2010.



Processing speed does accelerate much faster on a classical computing cluster compared to cloud computing (Cavet et al. 2012)

Example: Archival expertise CINES

Les services d’archivage au CINES



PAC

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

Assurance qualité
OAIS
Compétences archivistiques
Expertise formats
Processus métier
Gestion des risques

ISAAC

EUDAT

→ Archivage intermédiaire de données scientifiques

→ Archivage de données scientifiques pour des communautés européennes structurées