Scientific Data Preservation

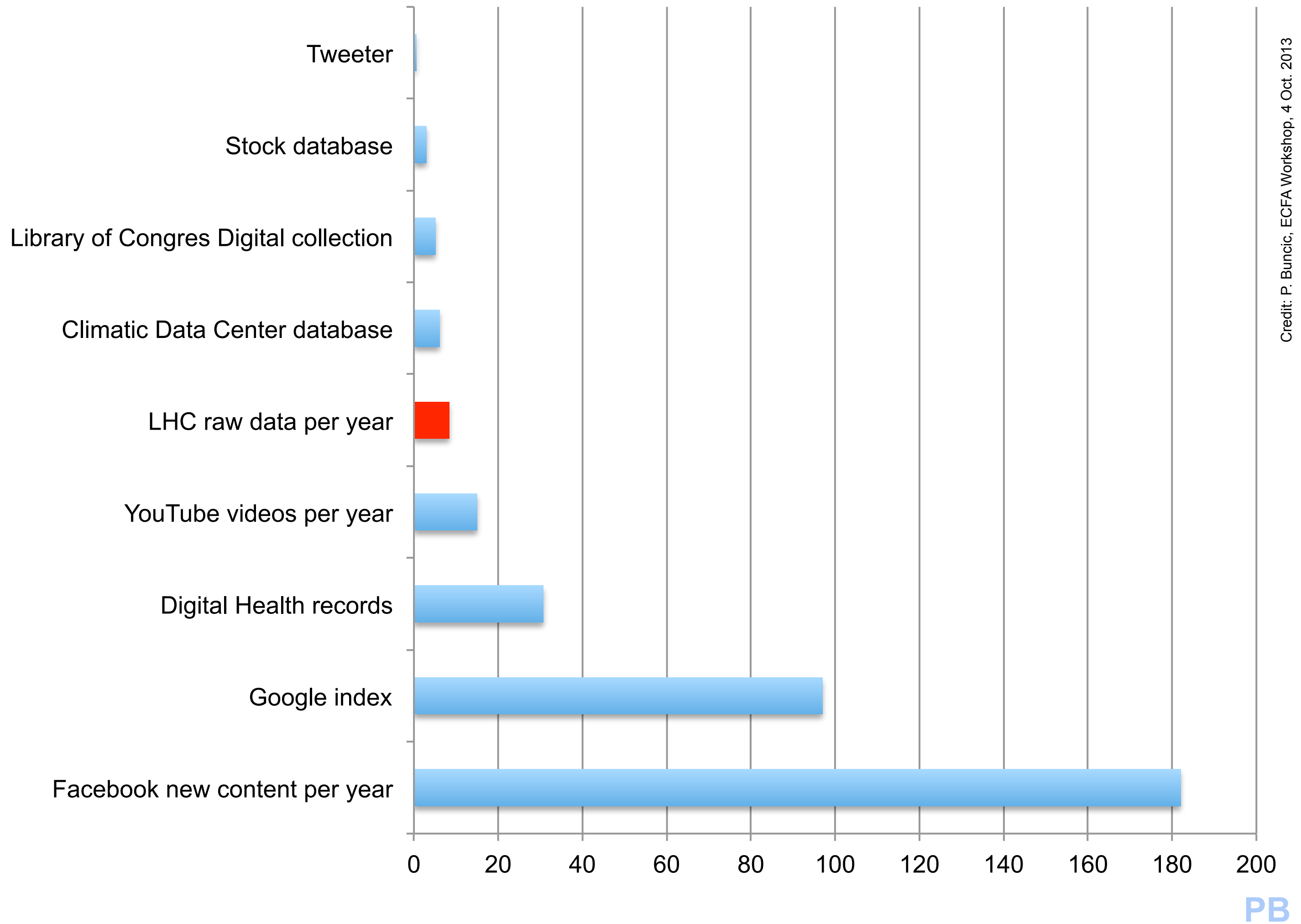
C. Diaconu

Centre de Physique des Particules de Marseille

CPPM/IN2P3/CNRS

Data Big Bang

Data collection accellerates



Digital data are fragile

- Storage capacity is physically exceeded
- Unattended/orpahned data vanishes quickly

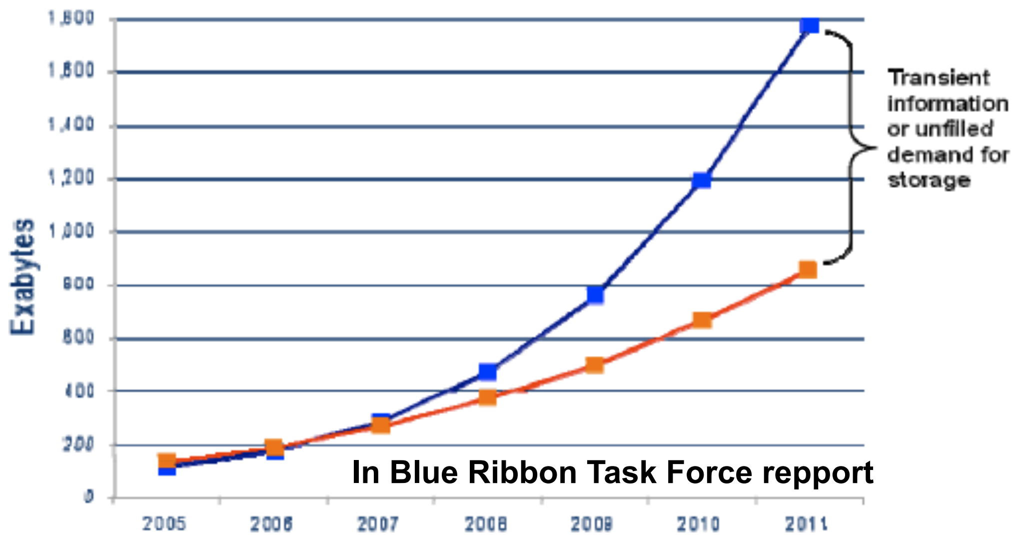


FIGURE 1.3: **Information and Storage**
Source: J. Gantz January 2008 (revised). Used with permission.

Models of data preservation and acces

- Collaborations address this issue in a generic way
  - e.g. Blue Ribbon, APA, DPC, eSciDir, RDA …

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

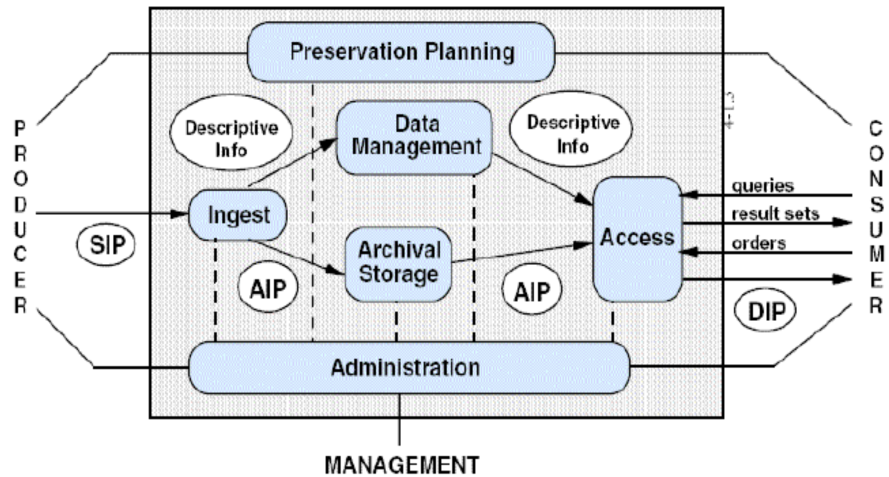
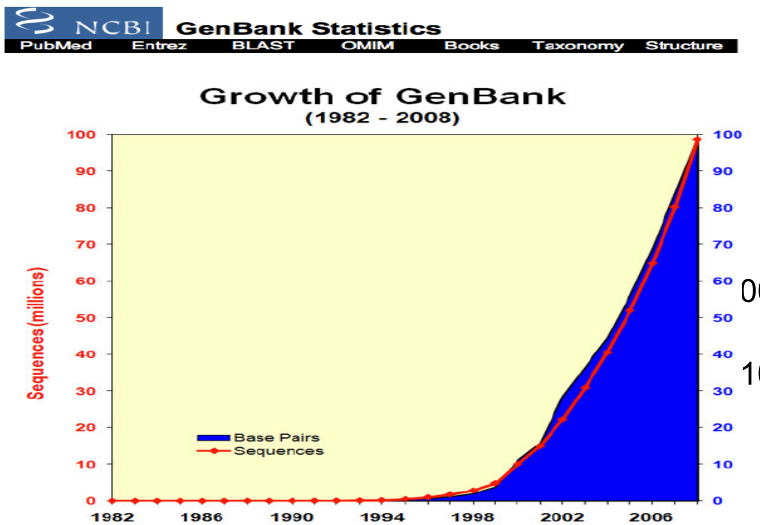


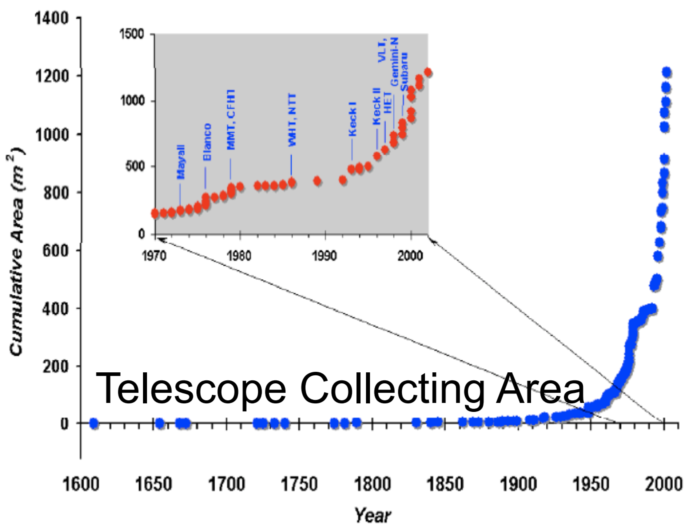
FIGURE 2.1: **The OAIS Reference Model**
http://public.ccsds.org/publications/archive/650x0b1.pdf, Page 4-1.
Source: Consultative Committee for Space Data Systems January 2002.

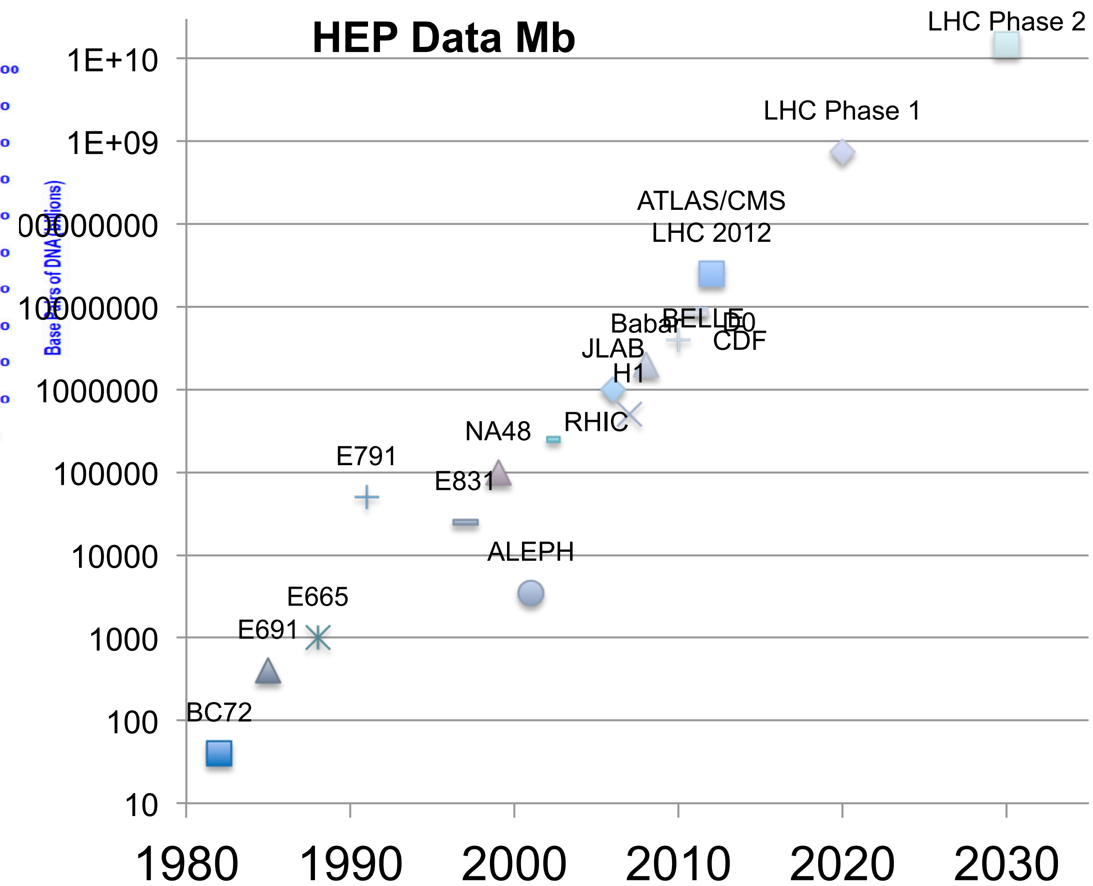
- Scientific Data is a major component of the ongoing efforts (complexity)

Big Scientific Data

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







Scientific Data

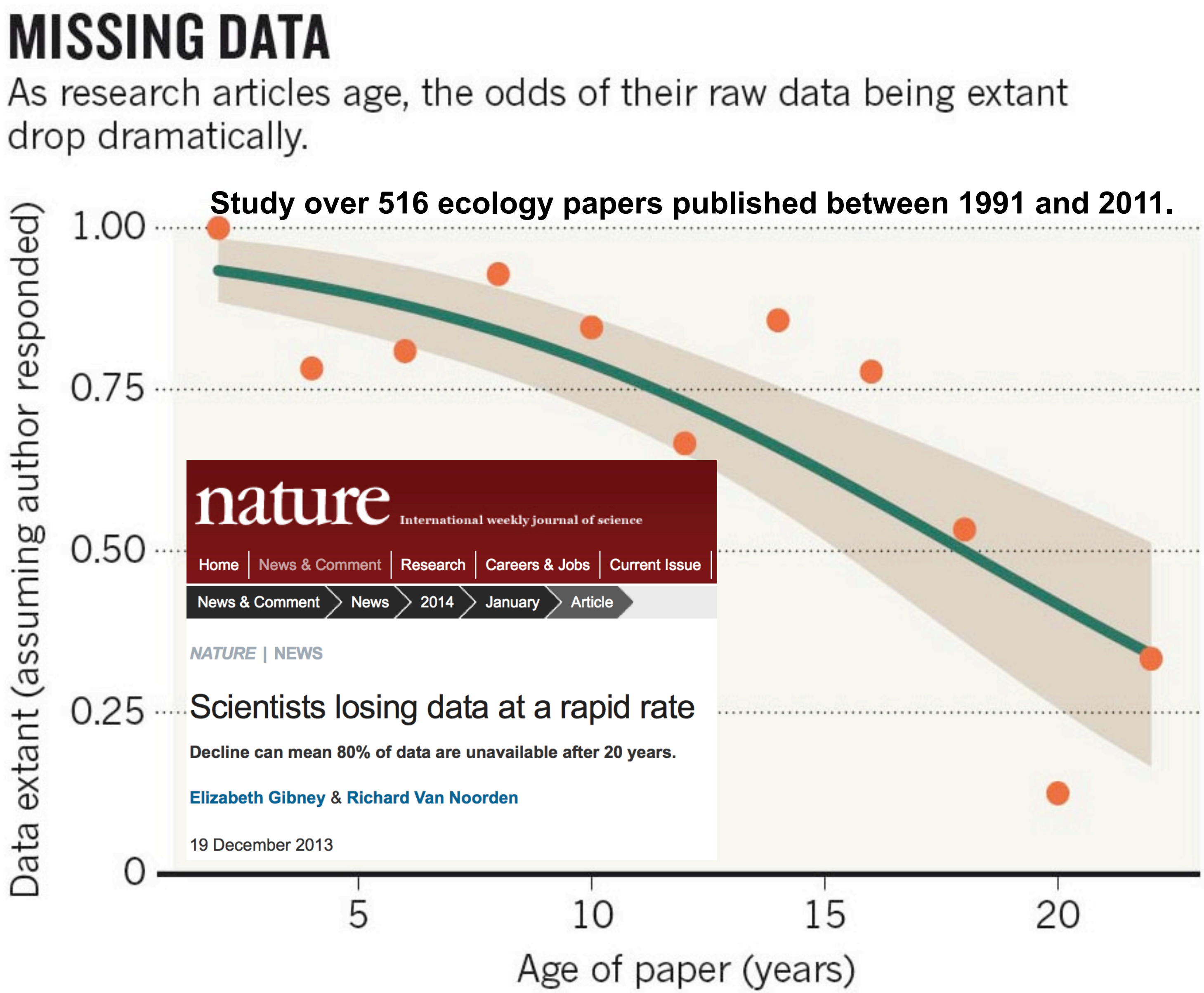
- Structured following a scientific plan
- Diverse sources
- Large and expensive projects
  - Not easy to repeat
- Contain unique knowledge
  - « Time stamped »
- Data Observatories
  - Contain more information than initially needed

Scientific endeavours: big and complex

- High energy physics projects
- Large Hadron Collider (27 Km, 13 TeV, 40MHz)
- 100Pb/year

Another project

Where is your data?



Preservation: where is the problem?

nature
International weekly journal of science

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

Archive > Volume 503 > Issue 7477 > News > Article

NATURE | NEWS

عربي

LHC plans for open data future

Researchers share results to keep them accessible.

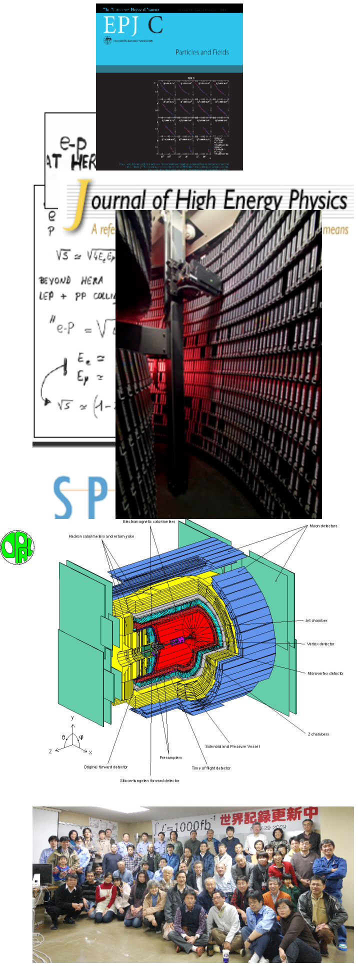
Elizabeth Gibney

26 November 2013

> **“When the LHC programme comes to an end, it will probably be the last data at this frontier for many years. We can’t afford to lose it.”**

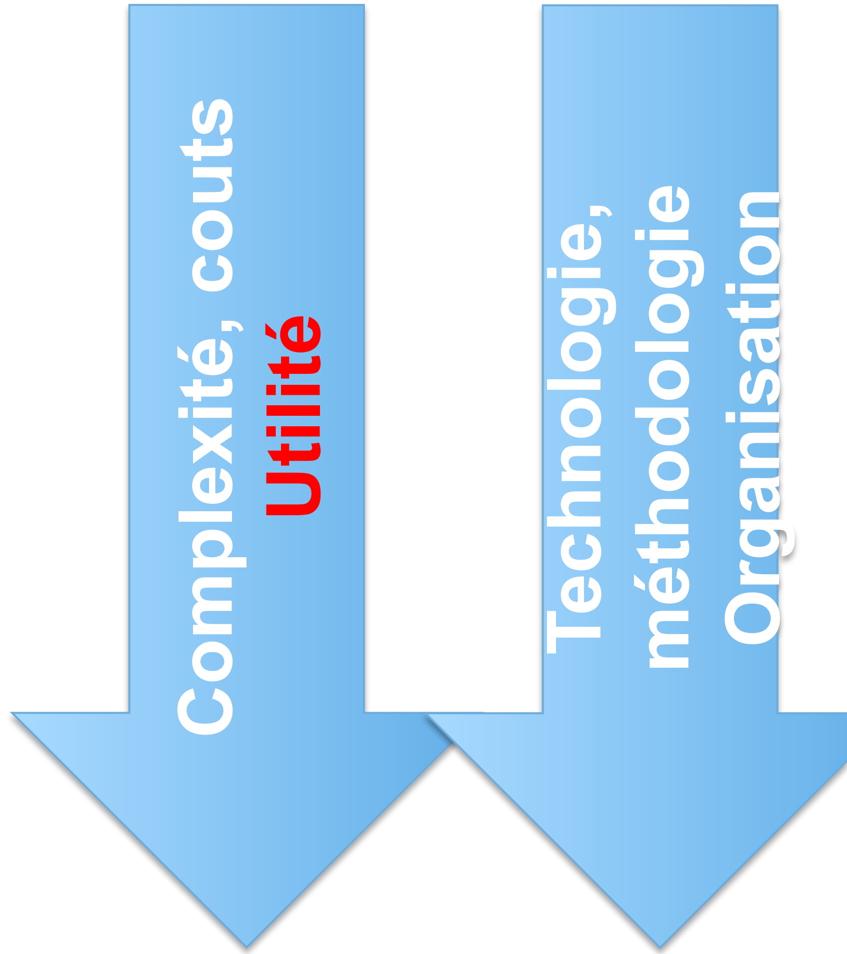
Storing the data is not a problem: hard drives are cheap and getting cheaper. The challenge is preserving knowledge that is less commonly stored — the software, algorithms and reference plots specific to each experiment. These often degrade or disappear with time, says Cristinel Diaconu of the Marseilles Centre for Particle Physics in France, who is chair of the international Data Preservation in Long Term Analysis in High Energy Physics (DPHEP) study group. He worries that if the data continue to be stored in their current state, physicists trying to decipher them in 10 years’ time will be unable to reconstruct the discovery of the Higgs boson. “When the LHC programme comes to an end, it will probably be the last data at this frontier for many years,” he says. “We can’t afford to lose it.”

Scientific Data: what is it?



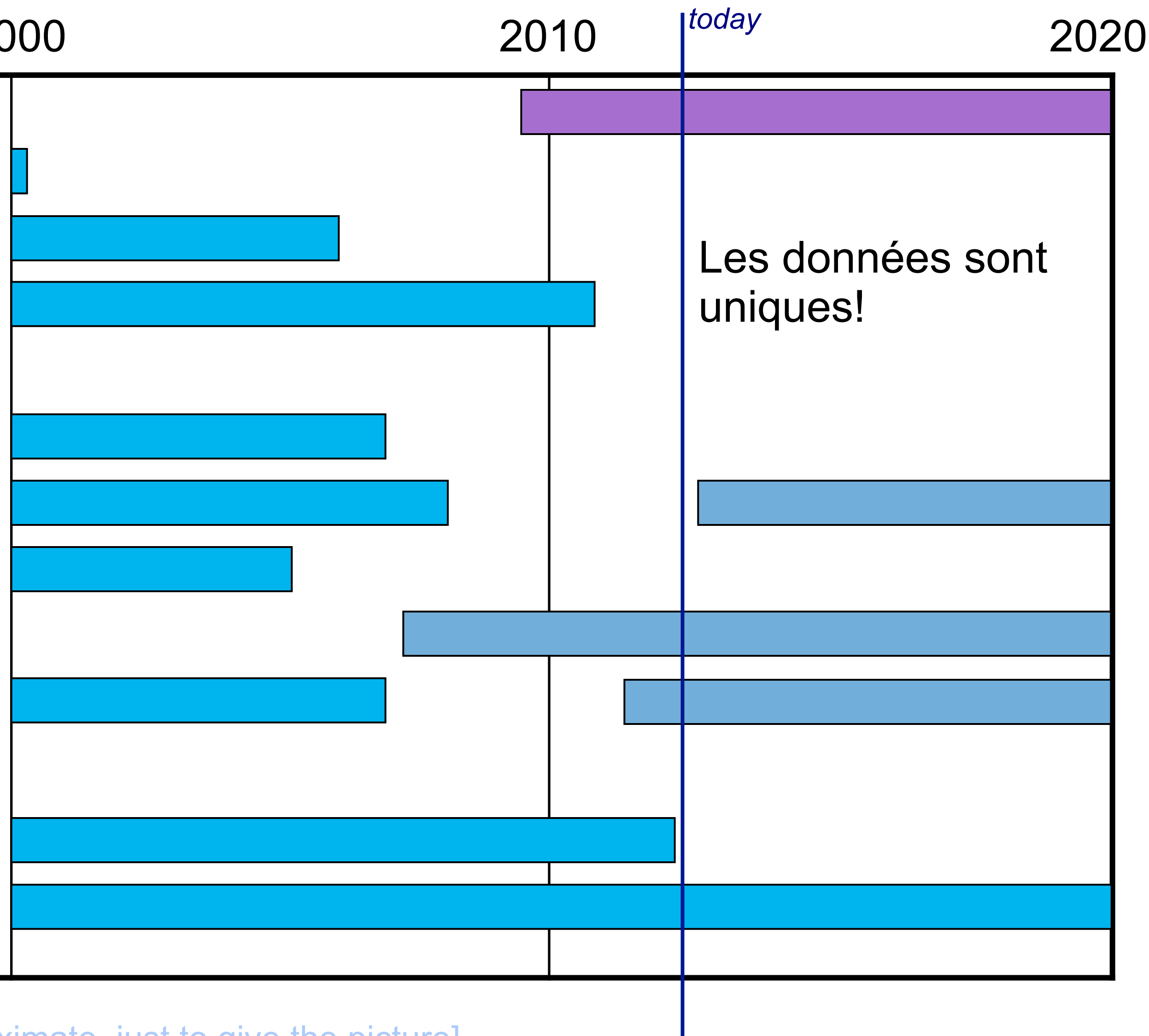
- Publications
- Documentation
- Raw
- Donées Processées
- Meta-données
- Workflows
- Software
- Diffuse knowledge

….more…



Exemple: HEP experiments in ± 10 ans

LHC	pp / ions	fb^{-1} ++
LEP	ee	0.9 fb^{-1}
HERA	ep	0.5 fb^{-1}
Tevatron	$p\bar{p}$	10 fb^{-1}
BaBar	ee	600 fb^{-1}
Belle	ee	1 ab^{-1} ++
CLEO C	ee	0.9 fb^{-1}
BES III	ee	fb^{-1} ++
KLOE	ee	1 fb^{-1} ++
RHIC	pp / ions	Multi-exp
SPS	Fixed target	Multi-exp



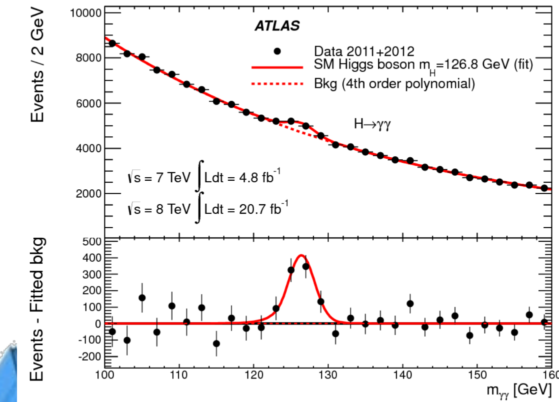
[not all programmes, dates are approximate, just to give the picture]

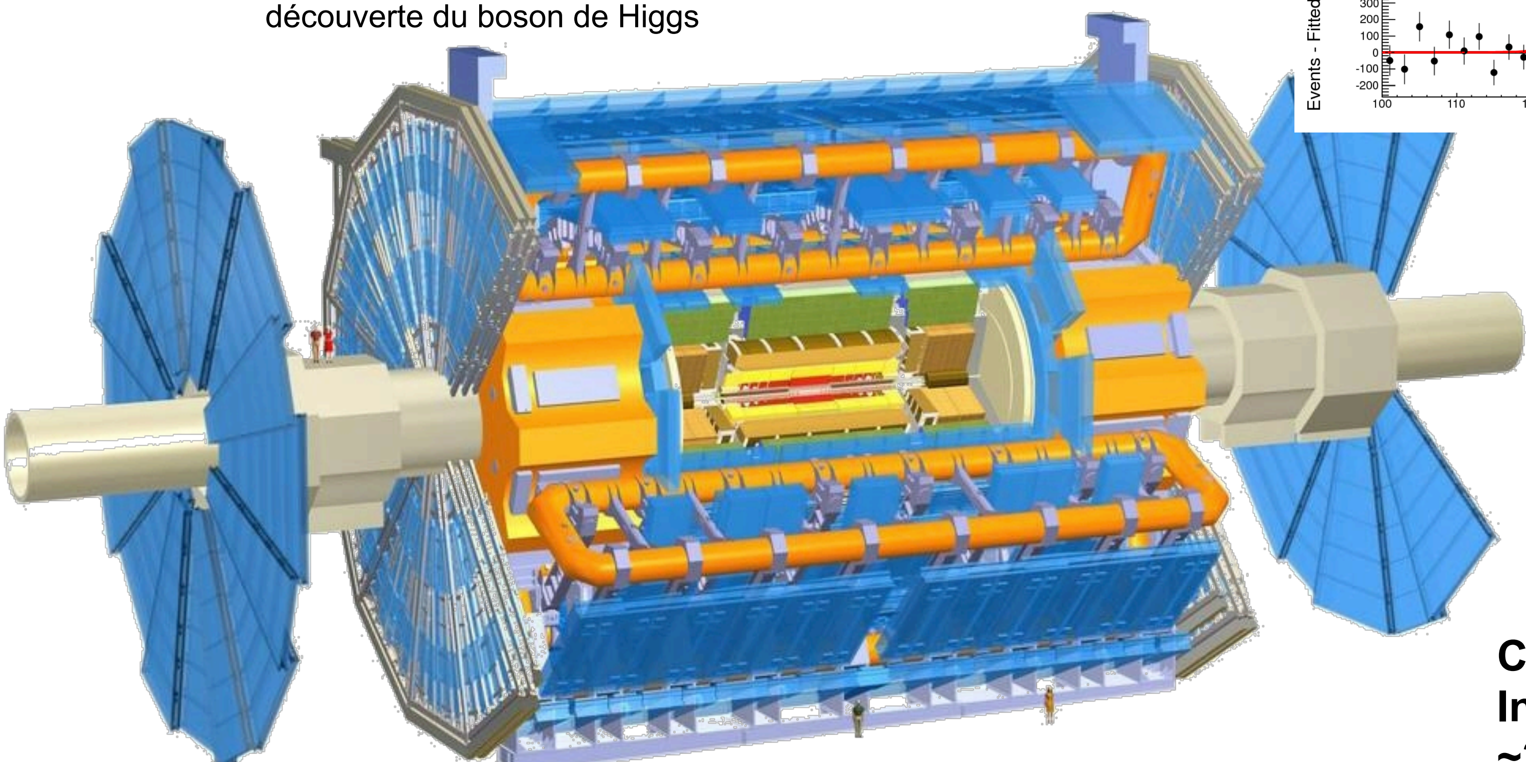
Complex experiments

- ATLAS: L'équivalent d'une camera avec 25Gpixels (avec une cinquantaine de technologies différentes) et 40 000 000 000 « photos » par seconde (100Pb)

- manips LHC:
  - 1000 articles scientifiques en 2 ans

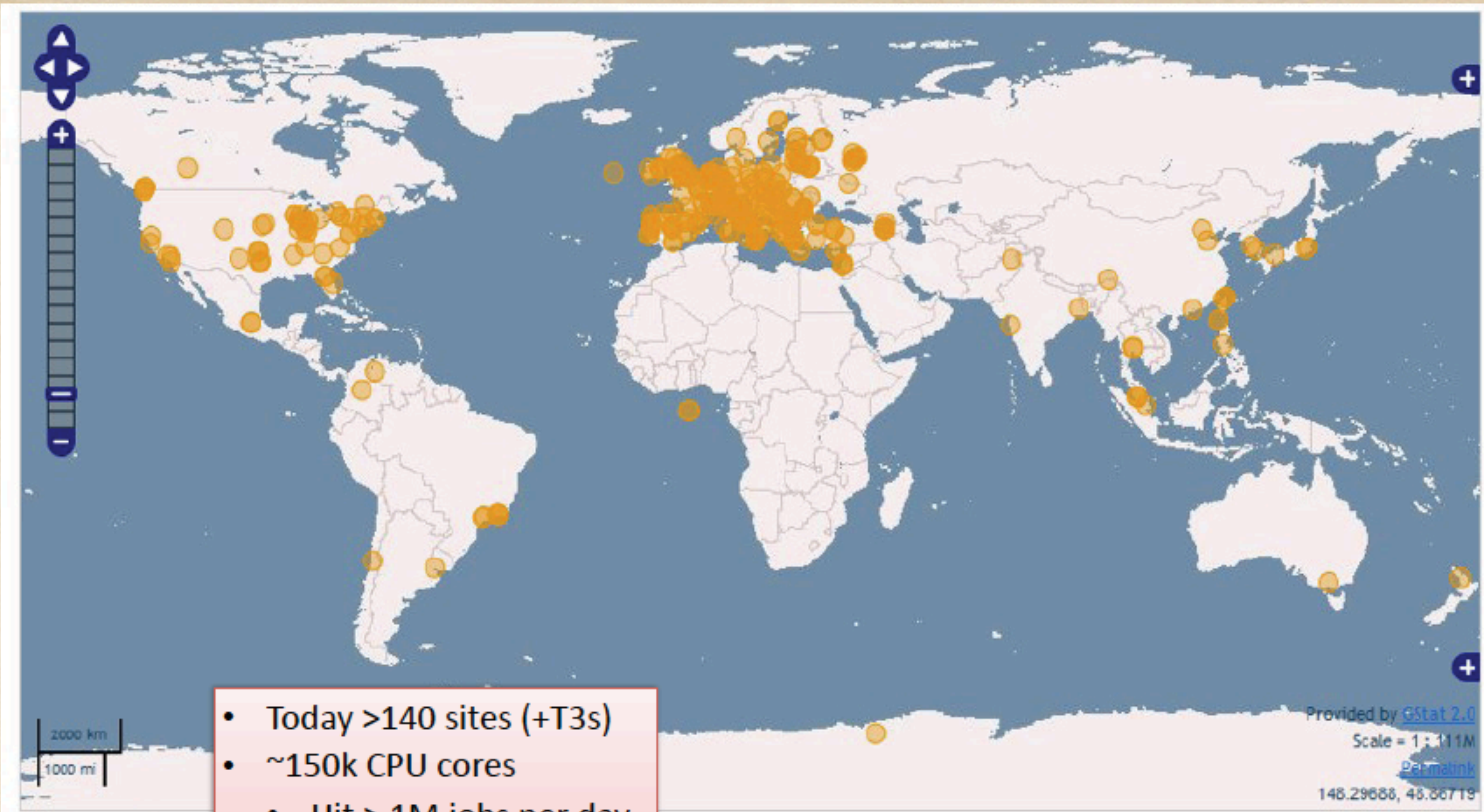
découverte du boson de Higgs





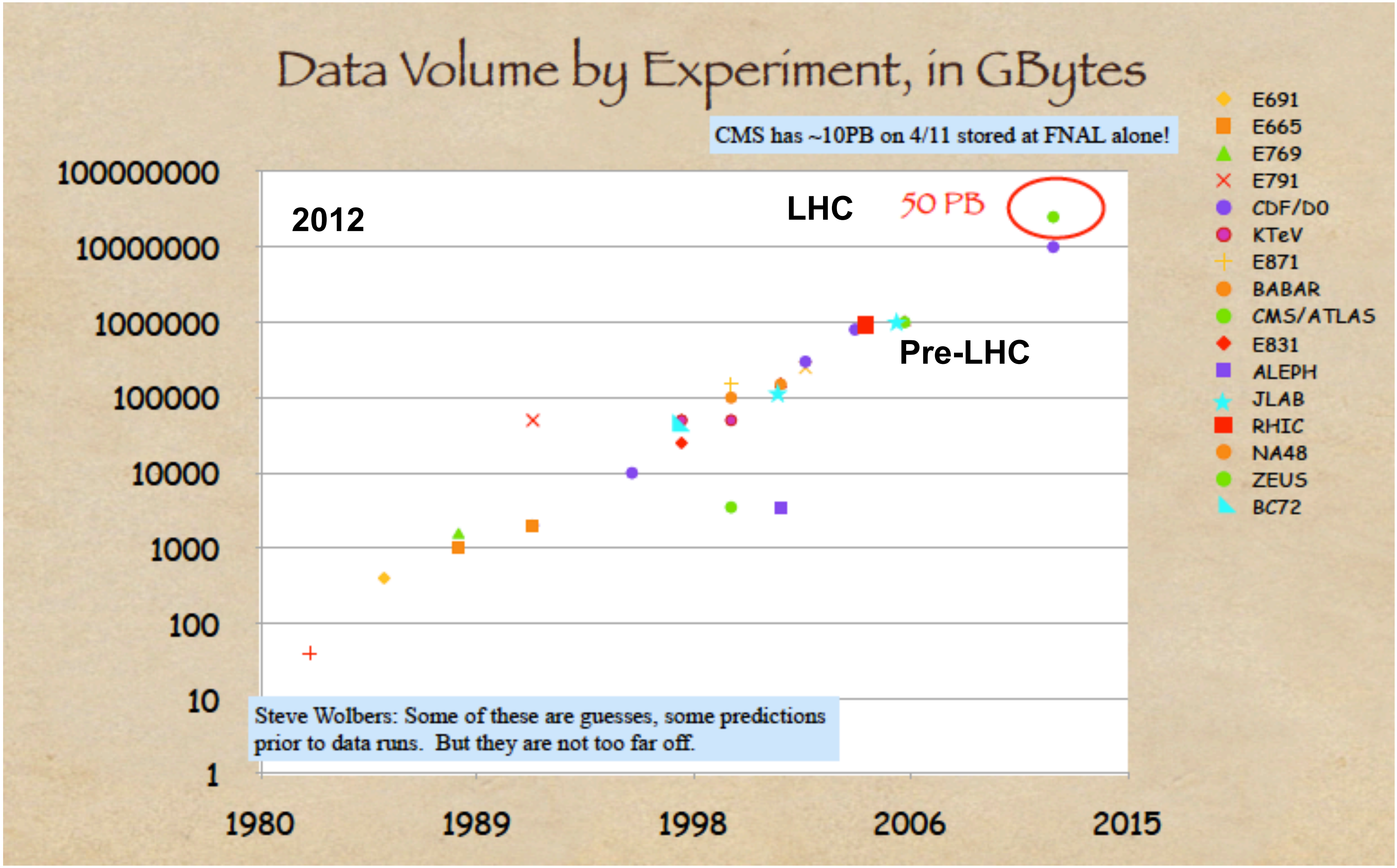
Collaboration Internationale ~3000 chercheurs

LHC computing

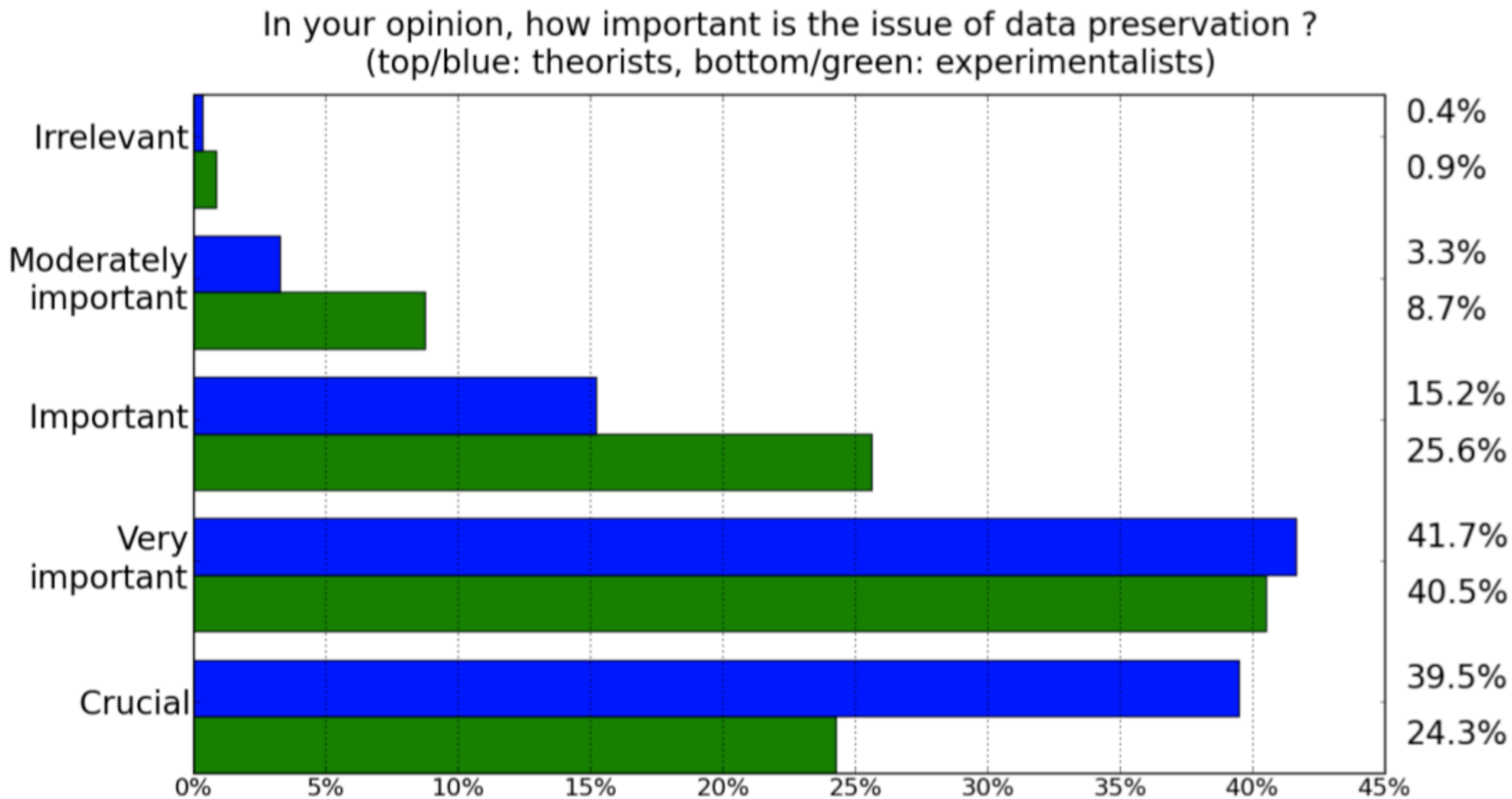


- Today >140 sites (+T3s)
- ~150k CPU cores
  - Hit > 1M jobs per day
- >50 PB disk

Data sets in time: 1PB -> 100PB->1EB



Should these data be preserved?



PARSE insight Permanent Access to the Records of Science in Europe

e-infrastructure

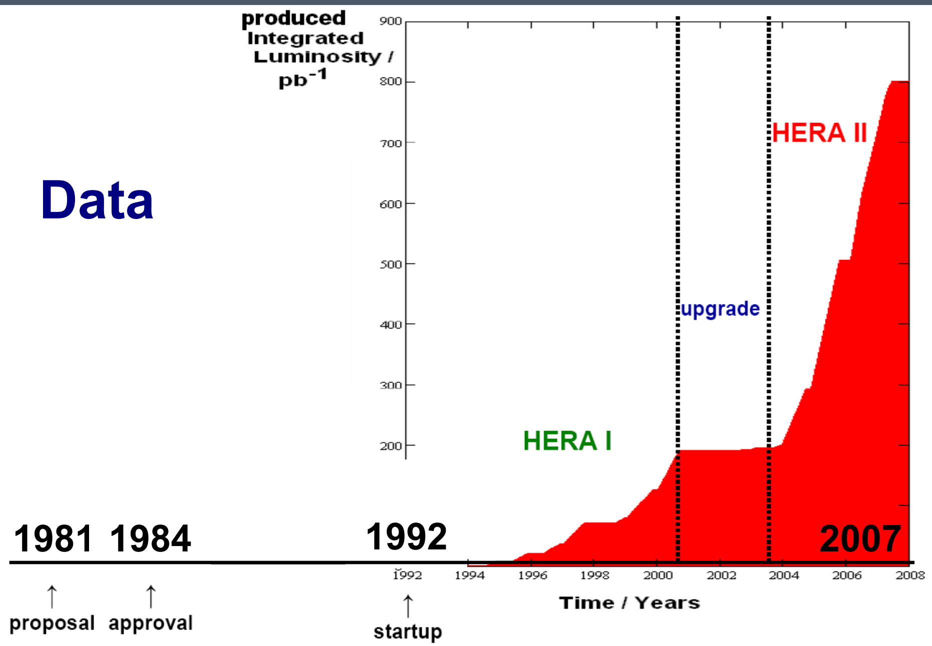
SEVENTH FRAMEWORK PROGRAMME

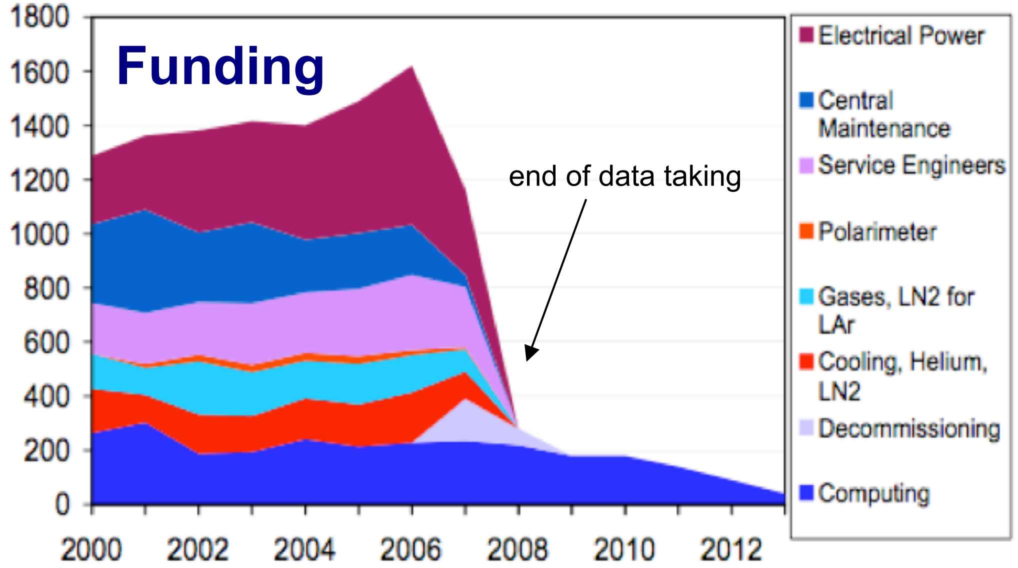
arXiv:0906.0485

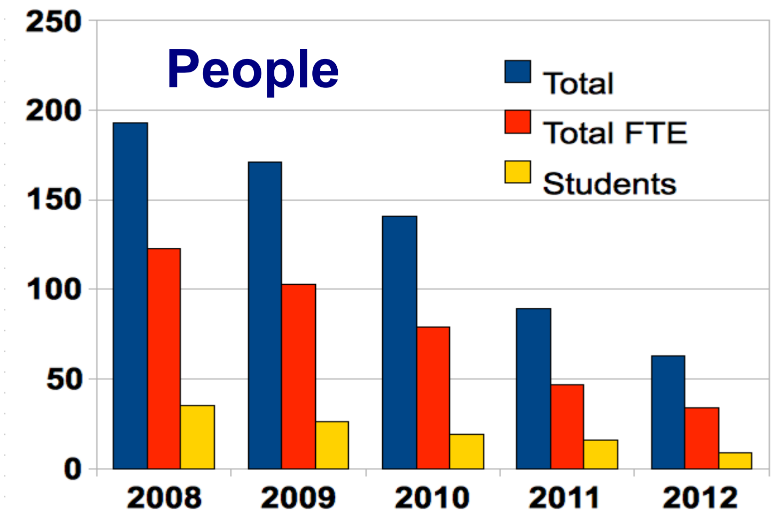
Why there may be a problem?

- End of the programs is difficult
  - Resources
  - Experts
- Upfront planning is essential

Data







The email you may receive one day (I did)

Dear Dr. Diaconu,

In the tape storage area we still have 4132 tapes of type 3840 containing HERA data.

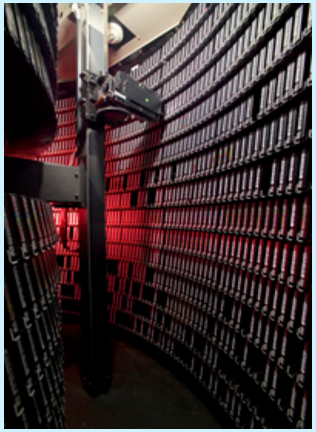
We do not have a functioning reading device anymore and **the storage area was polluted** recently, so **it is likely that the tapes are damaged**.

Would you like us to send you these tapes or should **we destroy them directly?**

Yours Sincerely,

Tape admin. service [a large computing centre]

What is “data”?



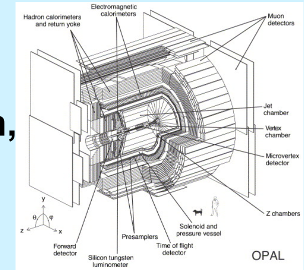
Digital information

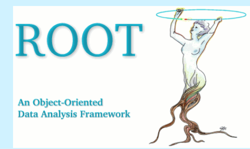
The data themselves, volume estimates for preservation data of the order of a few to 10 PB

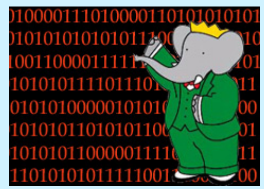
Other digital sources such as databases to also be considered

Software

Simulation, reconstruction, analysis, user, in addition to any external dependencies







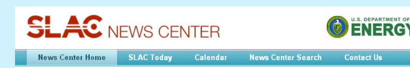
Software Licenses for LHC++ HEP-CCC: the licensing issues

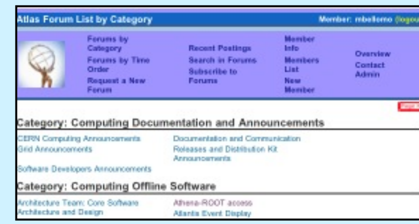
CERNLIB Access

- Access to the CERN Program Library is free of charge to all HEP users worldwide.
- Non-HEP academic and not-for-profit organizations: 1KSF/year

Meta information

Hyper-news, messages, wikis, user forums..





Publications

arXiv.org

iNSPIRE

HEPDATA: REACTION DATA Database

Durham University

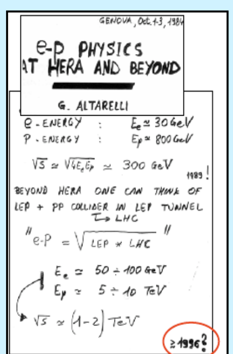
...containing numerical values of HEP scattering data such as total and differential cross sections, fragmentation functions, structure functions, and polarisation measurements, from a wide range of experiments. It is compiled by the Durham Database Group (UK) with help from the COMPAS group (Russia,) and is updated at regular intervals.

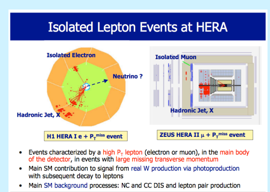
Journal of High Energy Physics

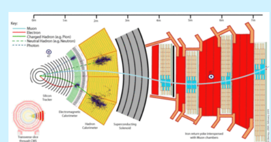
A refereed journal, written, run and distributed by electronic means

Documentation

Internal publications, notes, manuals, slides



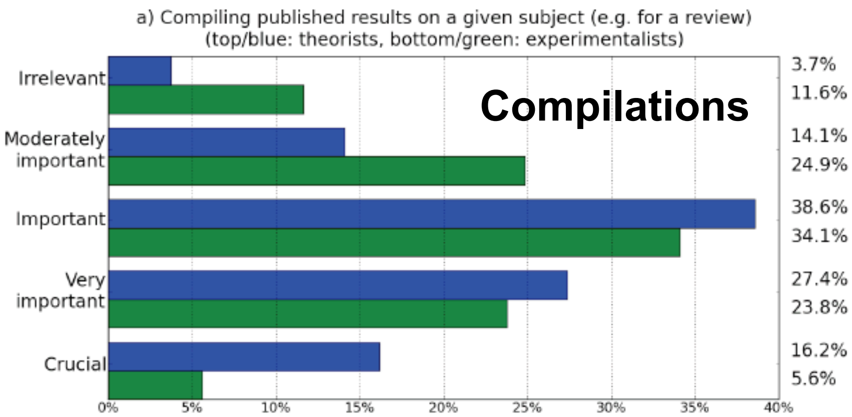


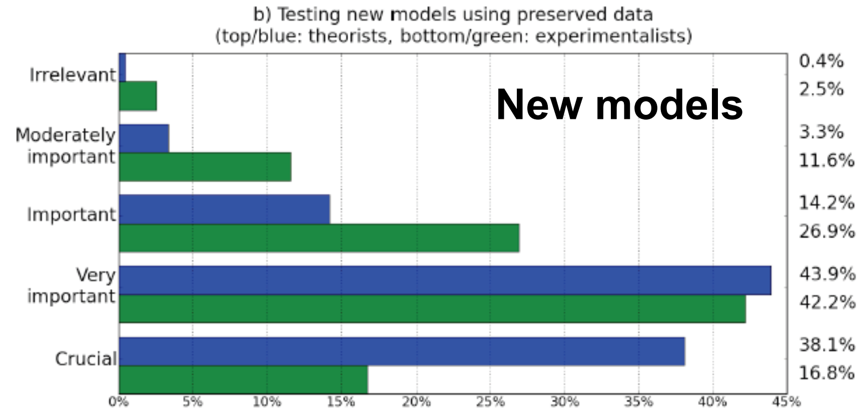


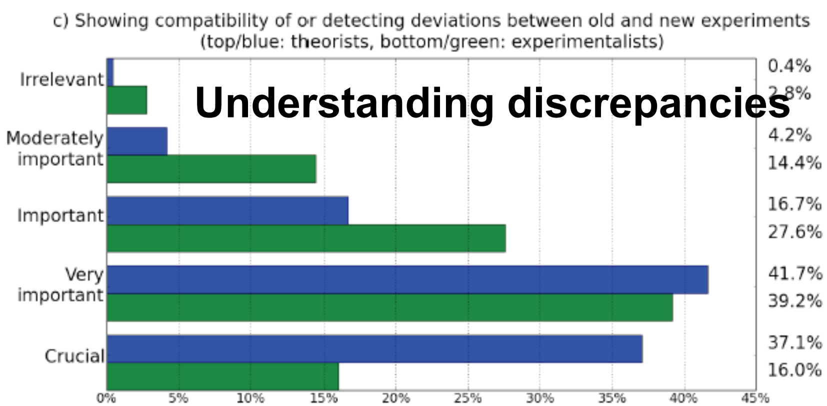
Expertise and people

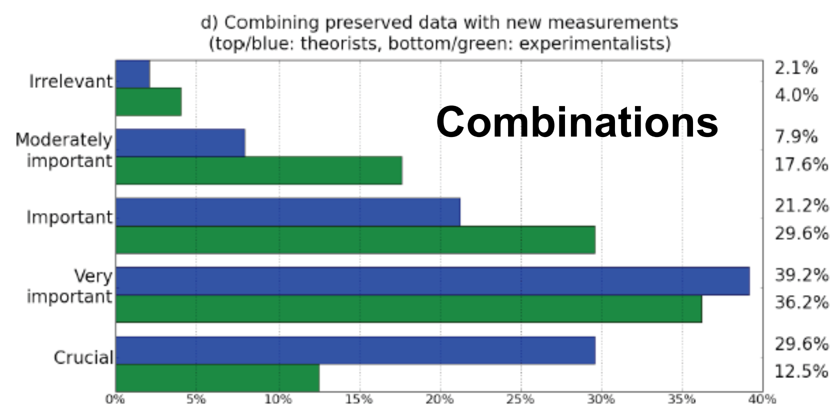
Data usage

Preserving HEP data is important for:









PARSE.Insight | Salvatore Mele | January 2009

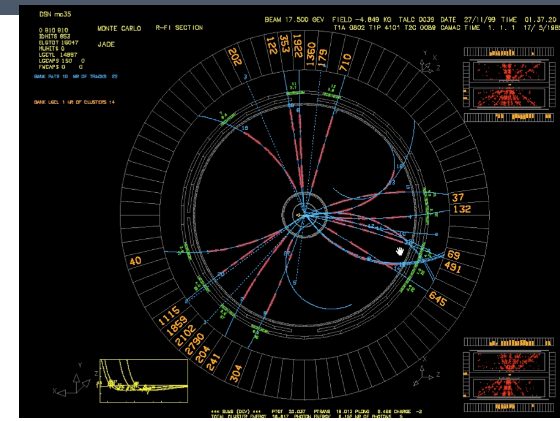
Rescued data used for fundamental results

- Experience JADE
  - Données sauvées par hasard

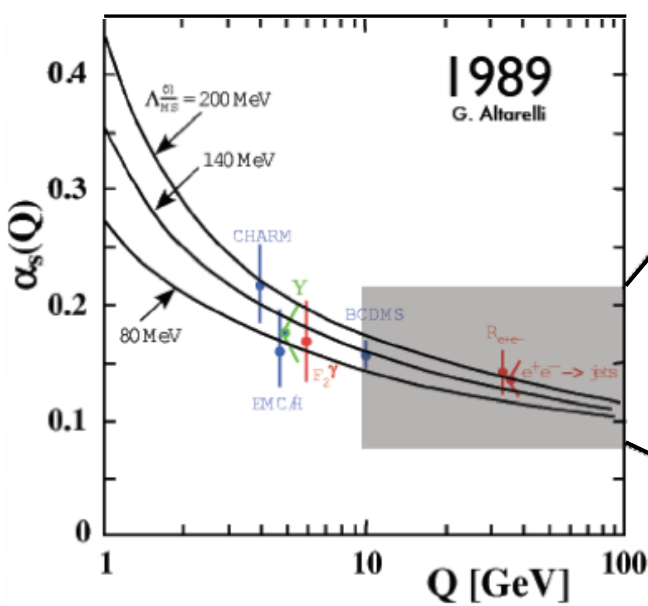
    Nom de code: “la valise”

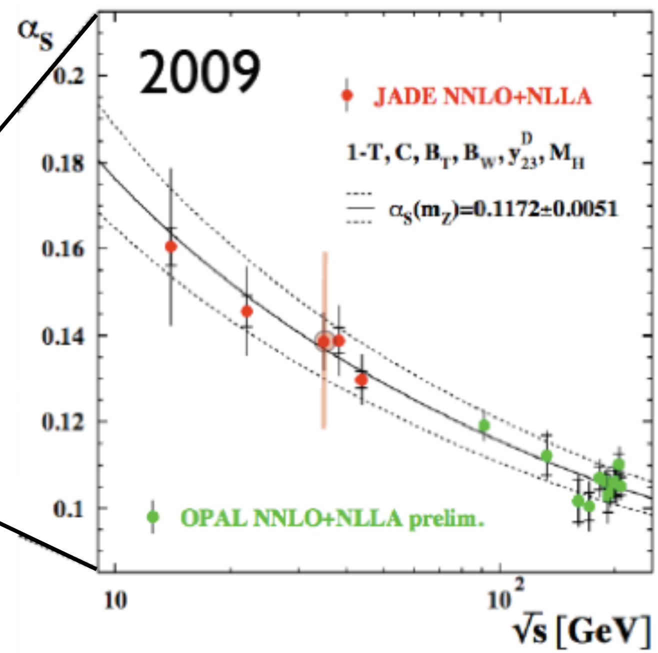
  - Ré-anlayse après 20 ans

10 publications

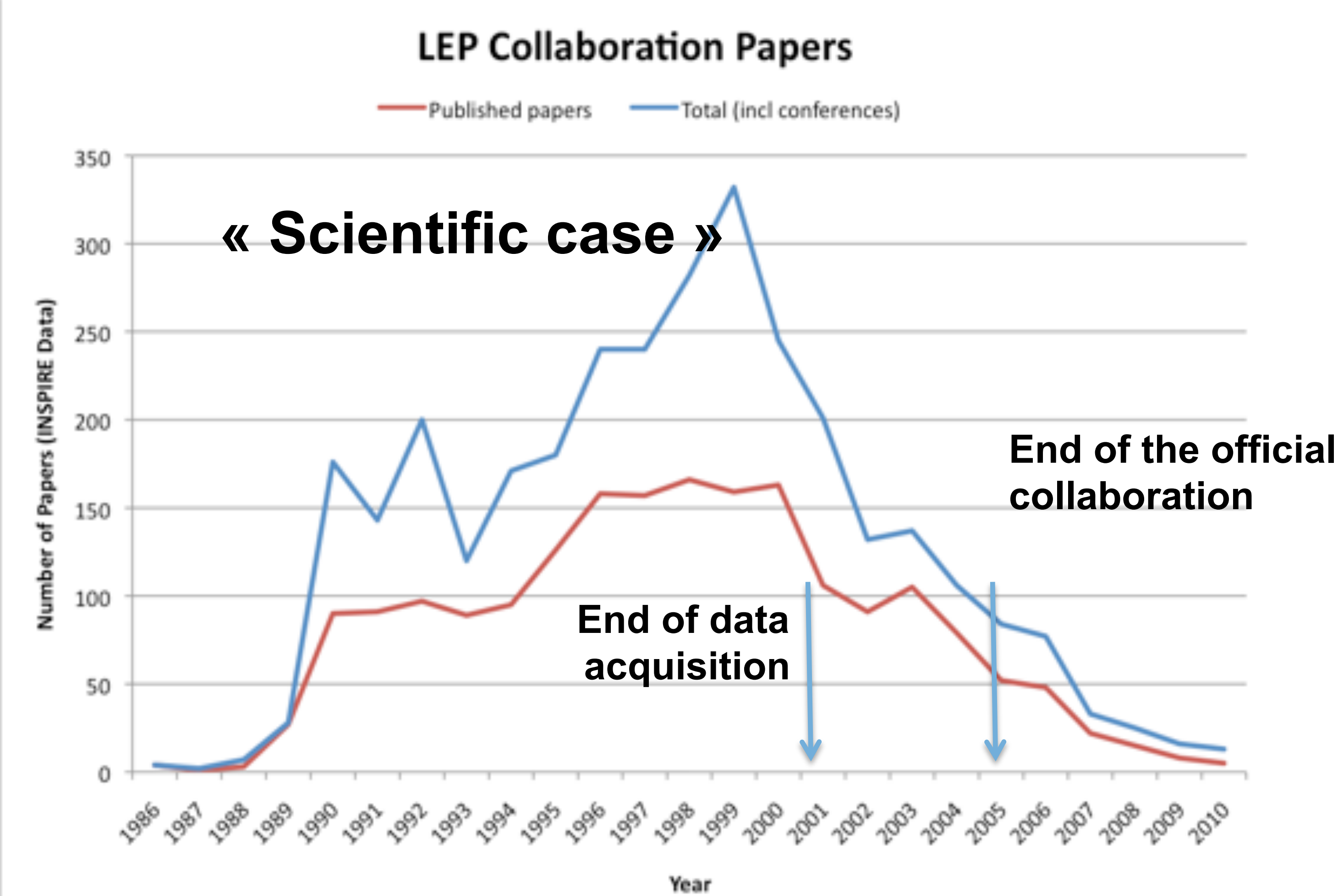


2011



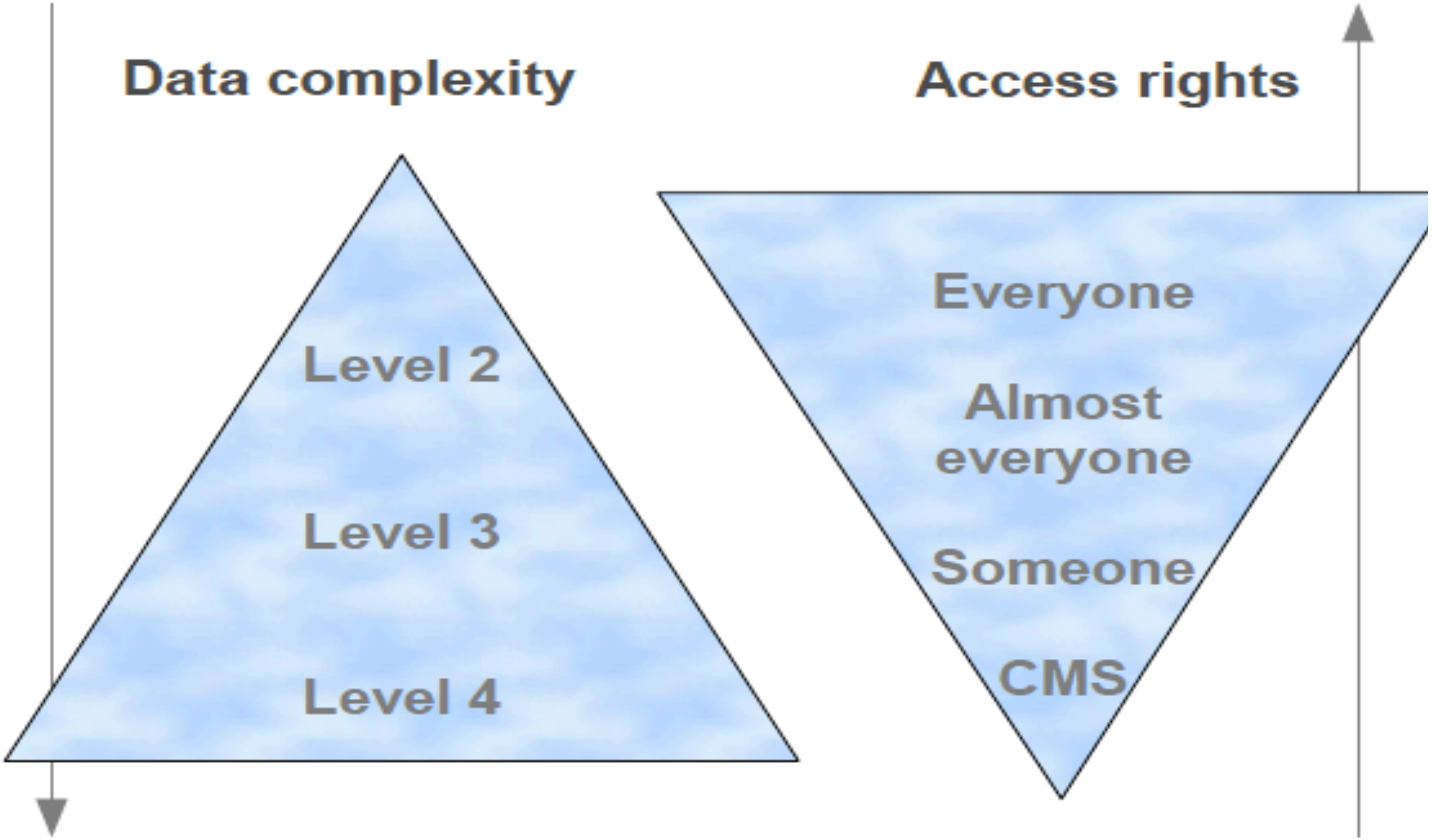


Long term publications yield



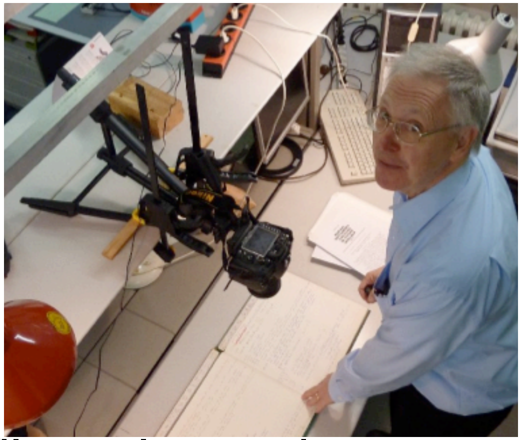
- Preserved data enables low cost science

Preservation complexity levels and access rights



Level 1: Documentation

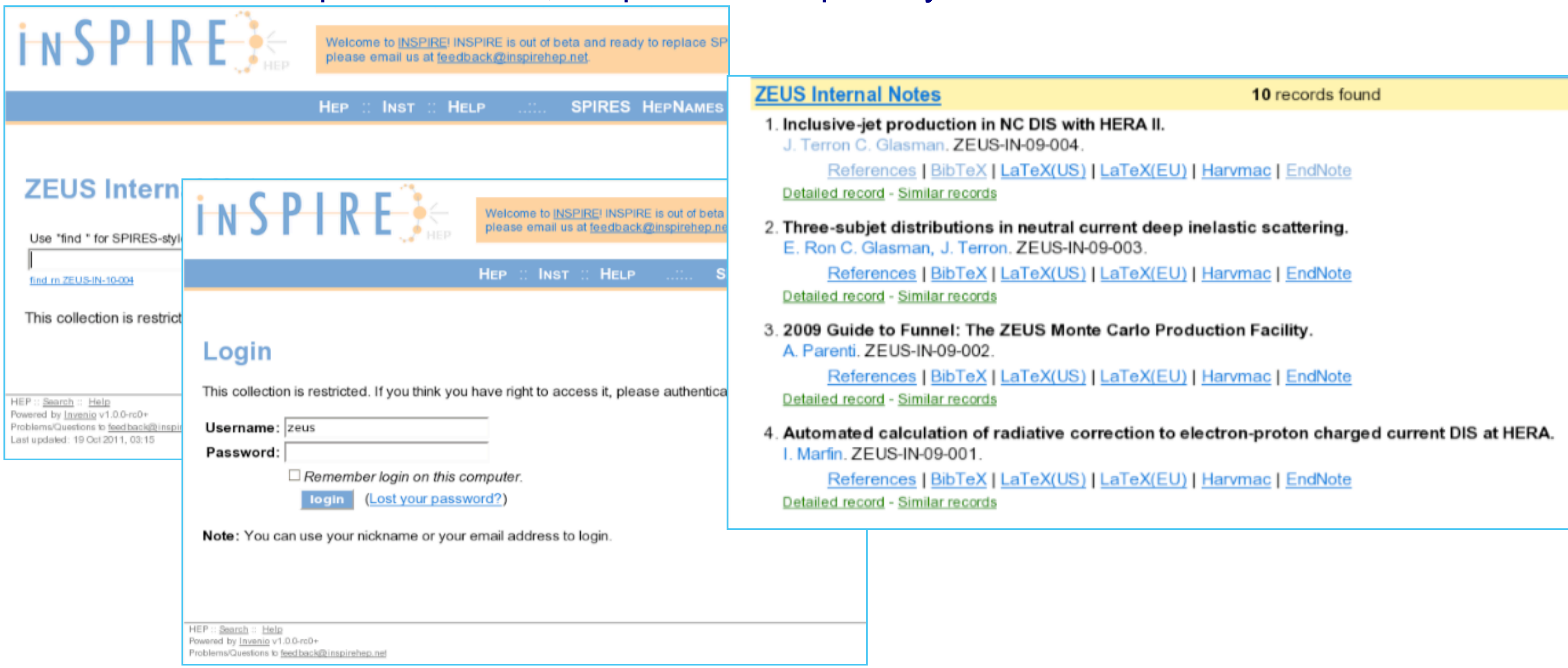
- Une tache considérable: des groupes de travail dédiées



- 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

Documentation projects with INSPIRE

- Internal notes now available on INSPIRE
  - Password protected now, simple to make publicly available in the future

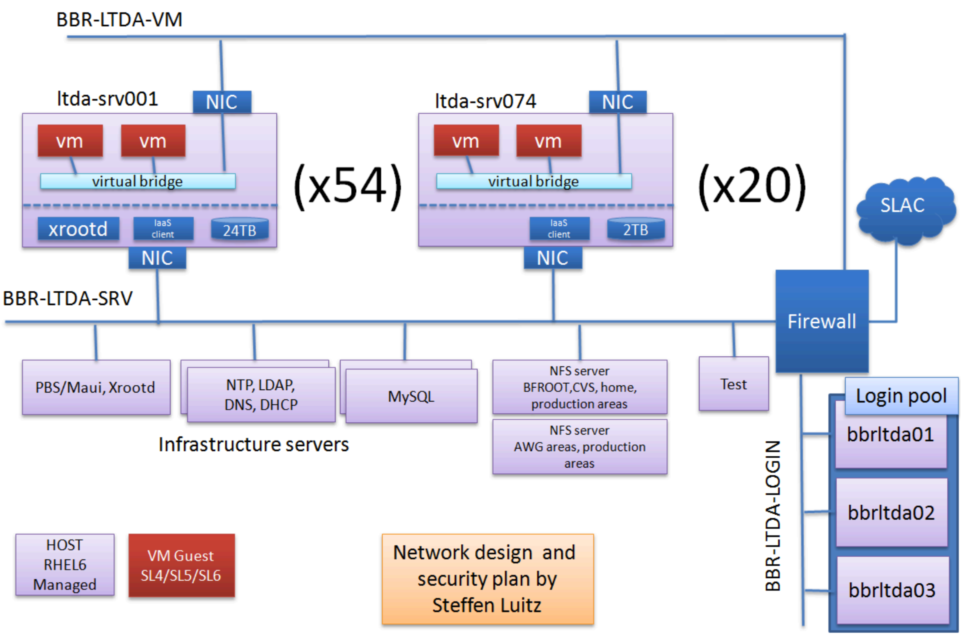


- The ingestion of other documents is under discussion, including theses, preliminary results, conference talks and proceedings, paper drafts, ...
- More on InSpire: reduced data?

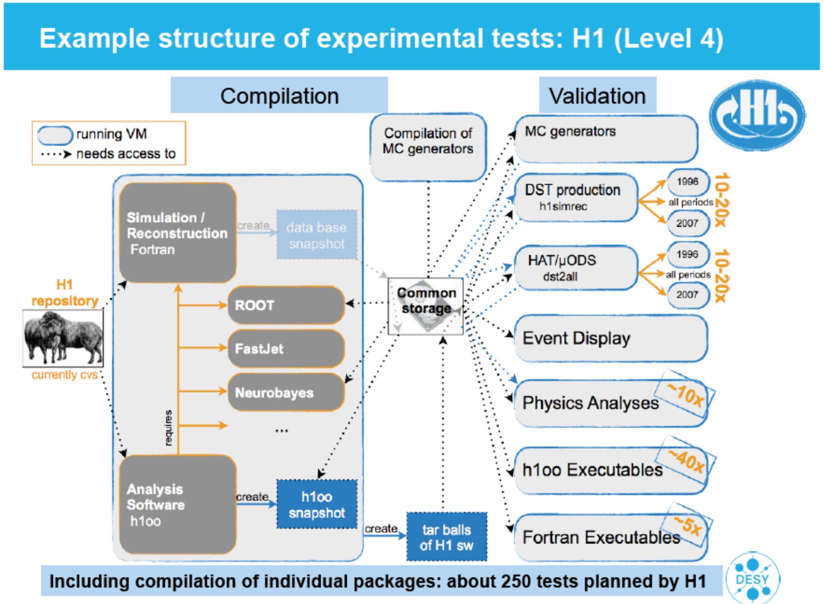
Exemples projets

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

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



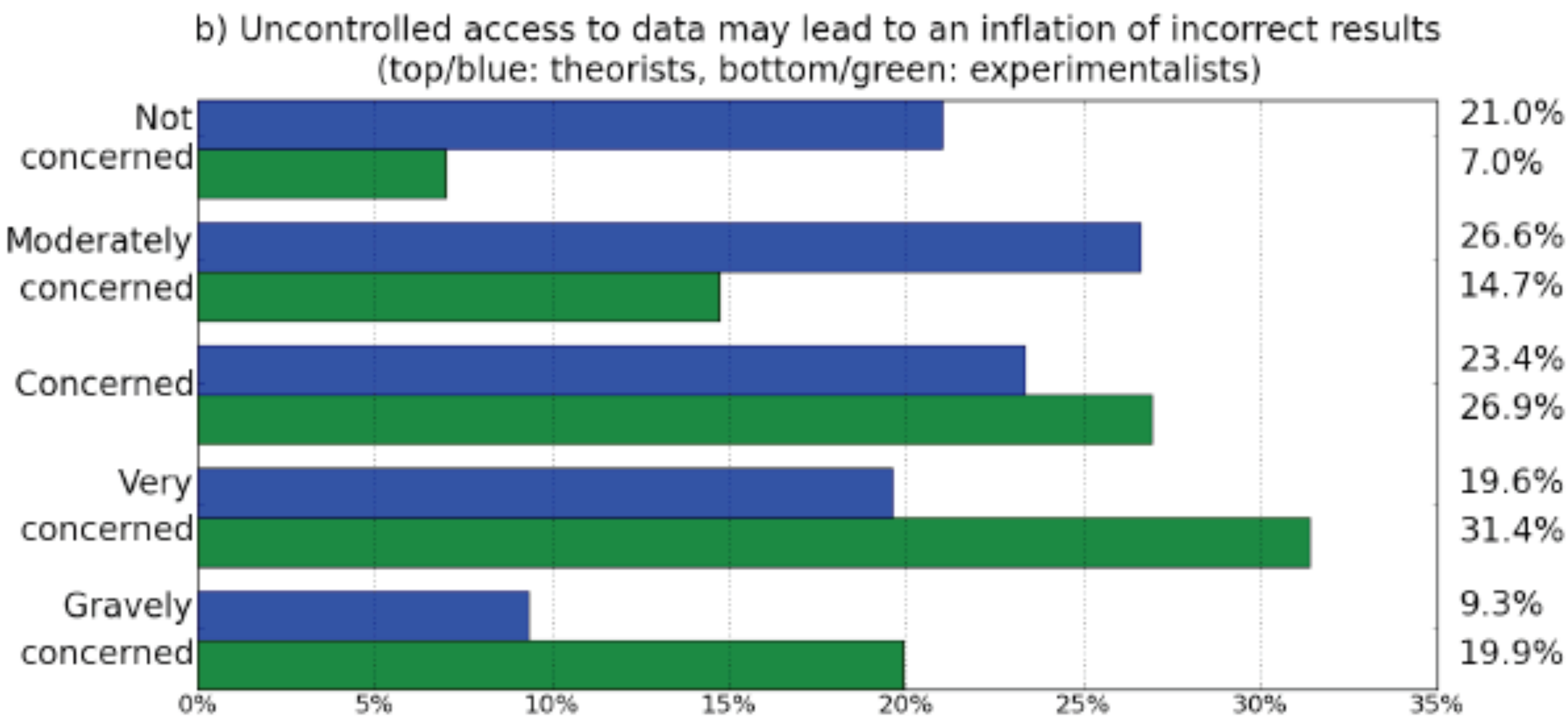
Système de préservation et migration Virtualisation, validation intensive (DESY, Hambourg, Allemagne)



Are complex preserved data « dangerous »?

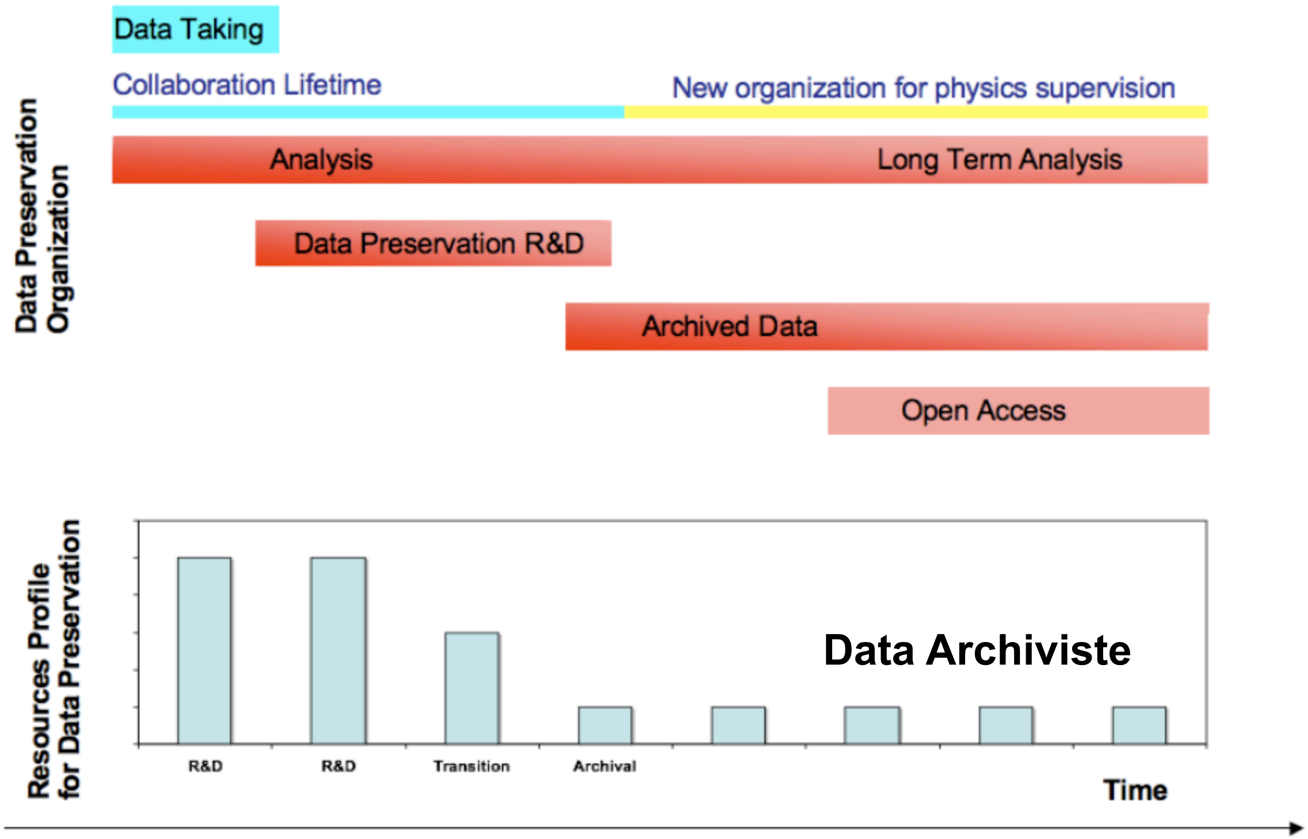
"Errors using inadequate data are much less than those using no data at all." Charles Babbage

Parse.insight



Governance issues are very important to support data usage

Long term organisation and economical models



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

DPHEP:International organisation

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

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

DPHEP Visibility

CERN Courier, May 2009

DATA PRESERVATION

Study group considers how to preserve data

For experimentalists in high-energy physics, the data are like treasure, but how can they be saved for the future? A study group is investigating data-preservation options.

High-energy-physics experiments collect data over long time periods, while the associated collaborations of experimentalists exploit these data to produce their physics publications. The scientific potential of an experiment is in principle defined and exhausted within the lifetime of such collaborations. However, the continuous improvement in areas of theory, experiment and simulation – as well as the advent of new ideas or unexpected discoveries – may reveal the need to re-analyse old data. Examples of such analyses already exist and they are likely to become more frequent in the future. As experimental complexity and the associated costs continue to increase, many present-day experiments, especially those based at colliders, will provide unique data sets that are unlikely to be improved upon in the short term. The close of the current decade will see the end of data-taking at several large experiments and scientists are now confronted with the question of how to preserve

A simulated event in the JADE detector, generated using a refined Monte Carlo program and reconstructed using revitalized software more than 10 years after the end of the experiment. (Courtesy Siggi Bethke.)

the complexity of the hardware and a more dynamic part closer to the analysis level. Data analysis is in most cases done in C++ using the ROOT analysis environment and is mainly performed on local computing farms. Monte Carlo simulation also uses a farm-based approach but it is striking to see how popular the Grid is for the mass-...ed events. The amount of data that should be ...is varies between 0.5 PB and 10 PB for each ...not huge by today's standards but nonetheless ...degree of preparation for long-term data varies ...but it is obvious that no preparation was fore-...of the programs; any conservation initiatives ...llel with the end of the data analysis.

Science

February 2011

Rescue of Old Data Offers Lesson for Particle Physicists

Old data tends to get forgotten as physicists move on to new and better machines.

naturenews

May 2011

symmetry

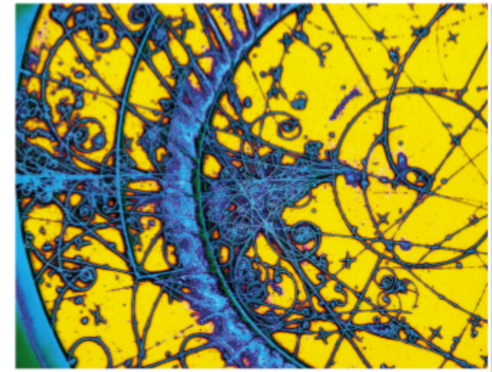
dimensions of particle physics

A joint Fermilab/SLAC publication

VOLUME 06 ISSUE 06 DECEMBER 09

Symmetry, December 2009

Wissenschaft



Die Hieroglyphen von morgen

An Beschleunigern sind immense Datenmengen entstanden – die Archivierung beginnt erst jetzt

Der Teilchenzoo

Berliner Zeitung and Frankfurter Rundschau, February 2010

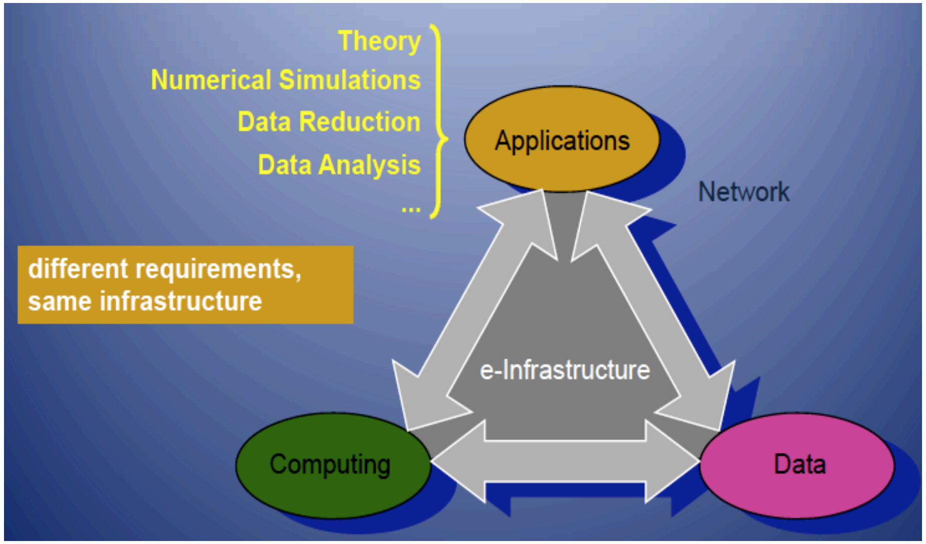
PDG particle data group

Data Preservation

- ICFA Study Group on Data Preservation and Long Term Analysis in High Energy Physics. High Energy Physics experiments initiated with this Study Group a common reflection on data persistency and long term analysis in order to get a common vision on these issues and create a multi-experiment dynamics for further reference:

  `https://www.dphep.org/`

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

F.Pasian

Initiatives in other fields

- Data preservation and in particular open access and data sharing are present in other fields such as:
  - Astrophysics, molecular biology, earth sciences, humanities and social science

Scientific Data preservation in a multidisciplinary approach

> Challenges:

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

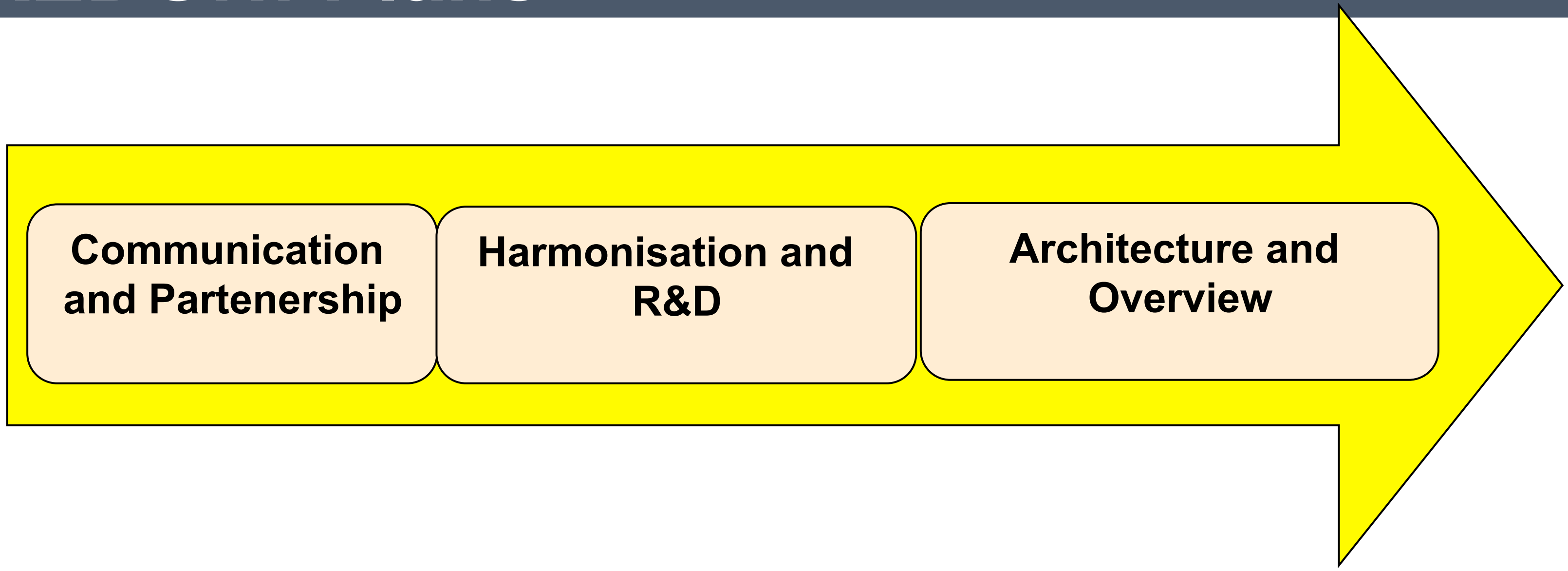
PREDON Consortium

PREDON

A project for scientific data preservation in France

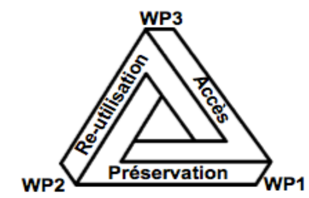
	Volume données	Complexité	Diversification des sources	Structuration au niveau international	Algorithmes et methodologies pour la preservation
IN2P3 HEP	+++	+++	+	++	+
INSU, IRD Astrophysics Earth Sciences	++	++	++	+++	++
CINES INS2I IT, Algorithms, workflows	+	++	+++	+	+++

PREDON: Plans



- **>** 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"

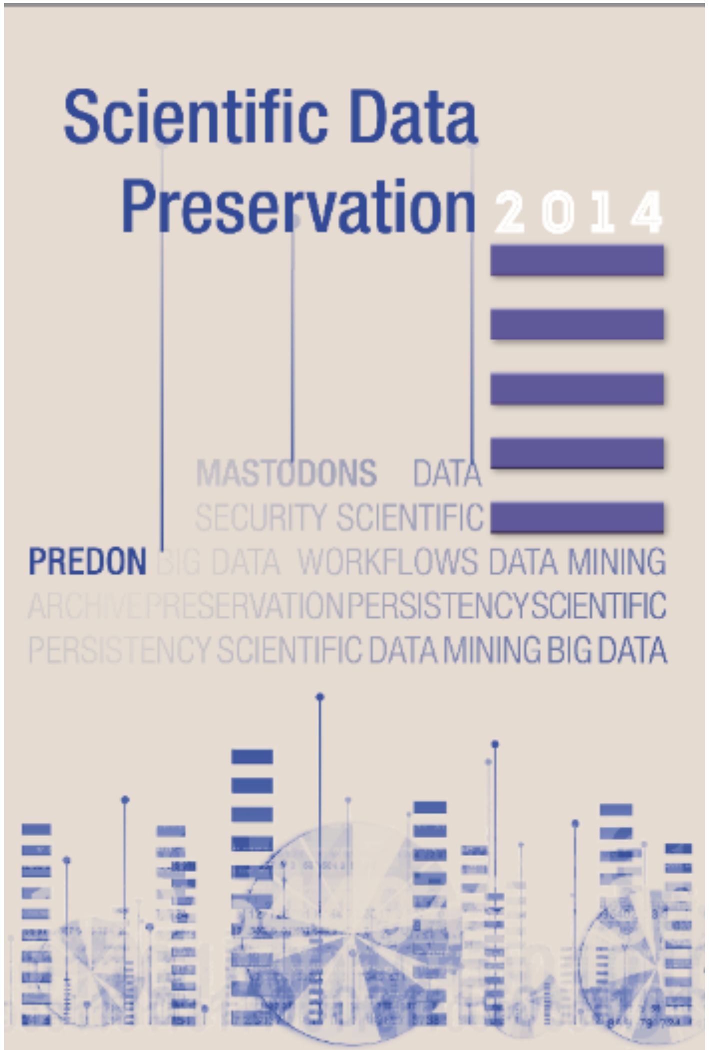
PREDON as a project



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*

Scientific data preservation: white paper

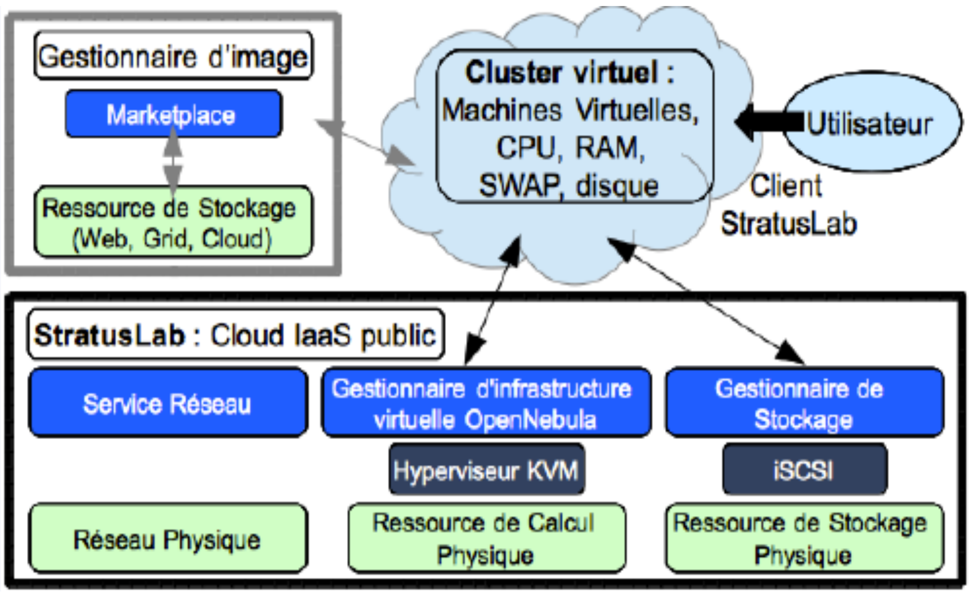
Available at htpp://predon.org



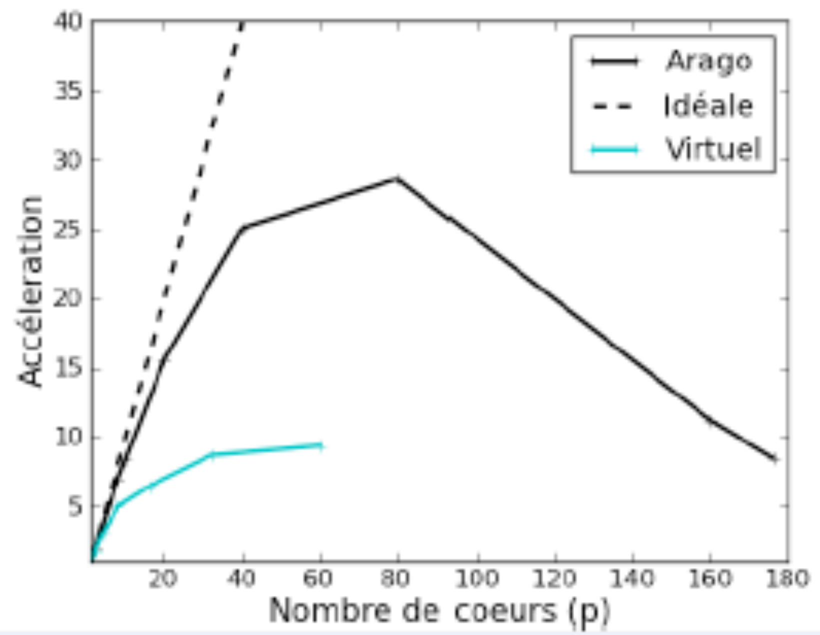
Exemple projet: Data processing & storage in the cloud

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



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

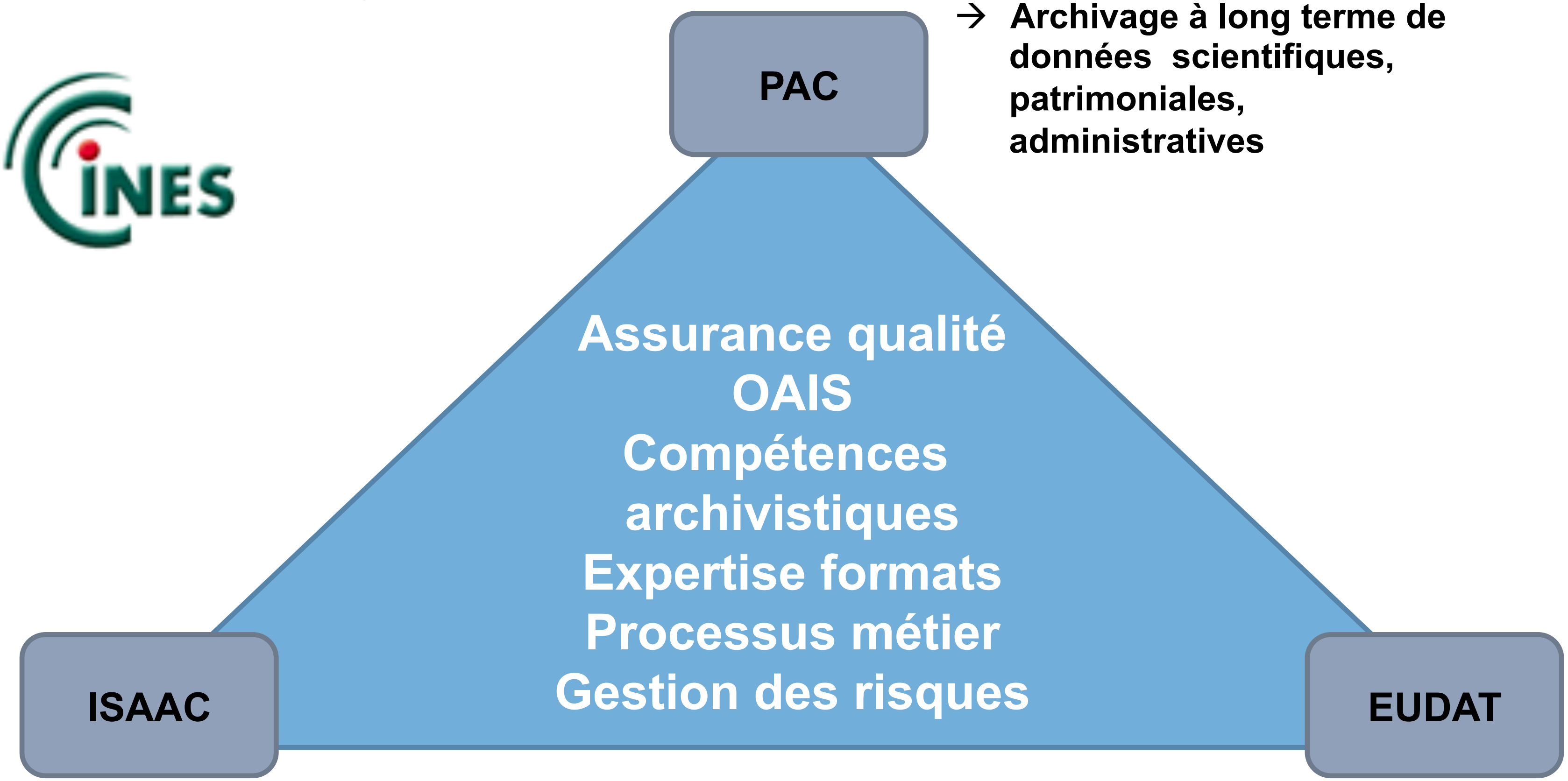


Processing speed does accelerate much faster on a classical computing cluster compared to cloud computing (Cavet et al. 2012)

Archival expertise CINES

Les services d'archivage au CINES

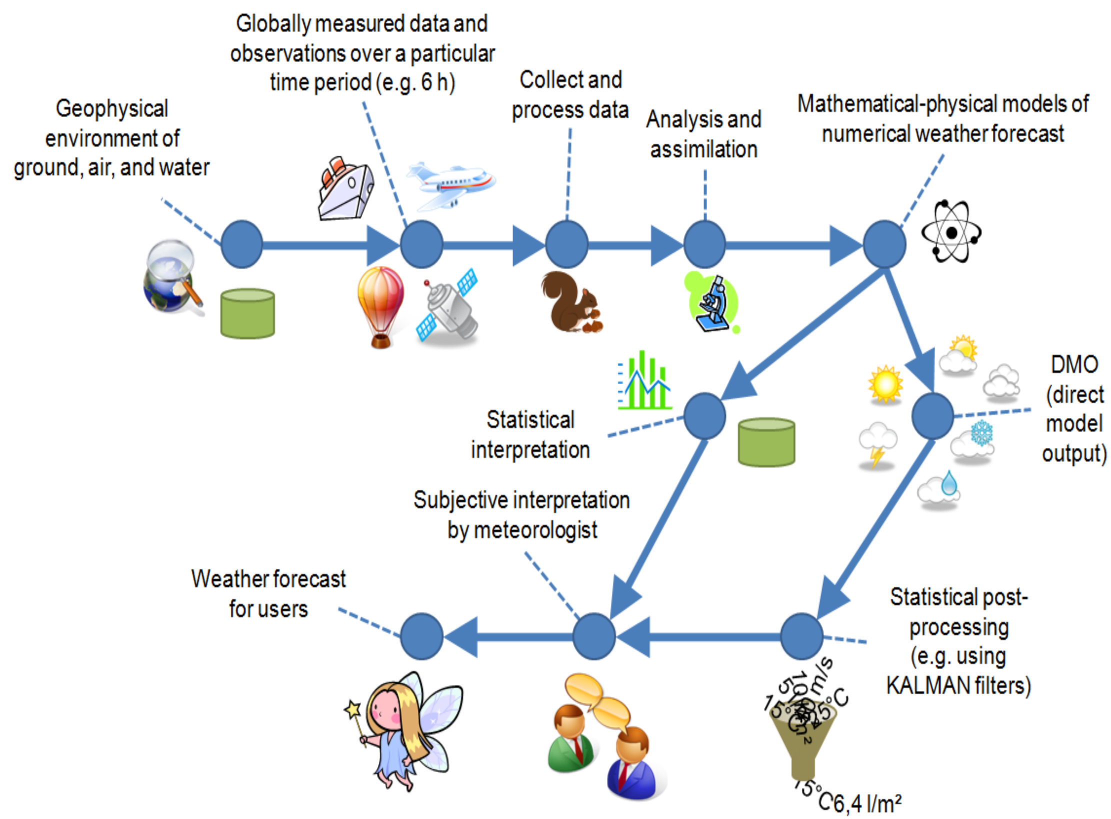
→ Archivage à long terme de données scientifiques, patrimoniales, administratives



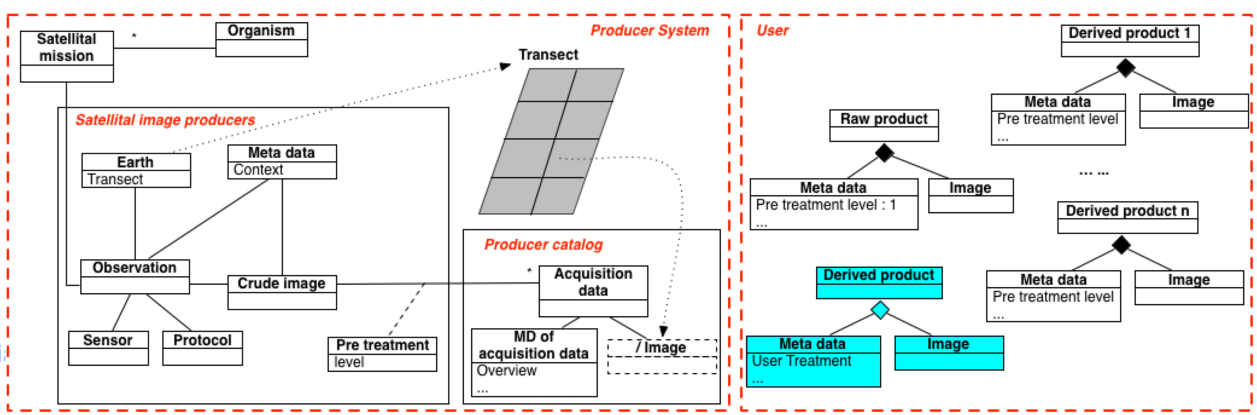
→ Archivage intermédiaire de données scientifiques

→ Archivage de données scientifiques pour des communautés européennes structurées

Workflows for data preservation



Rigorous approach emerges



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.

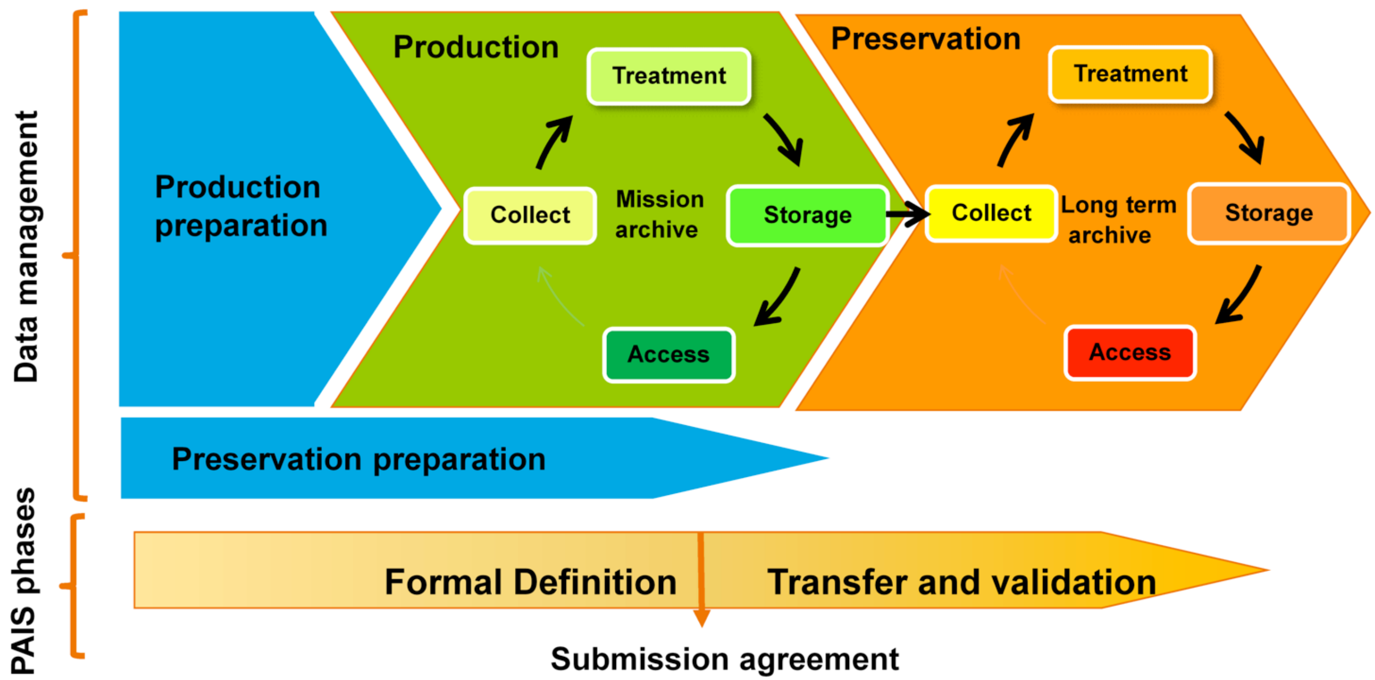
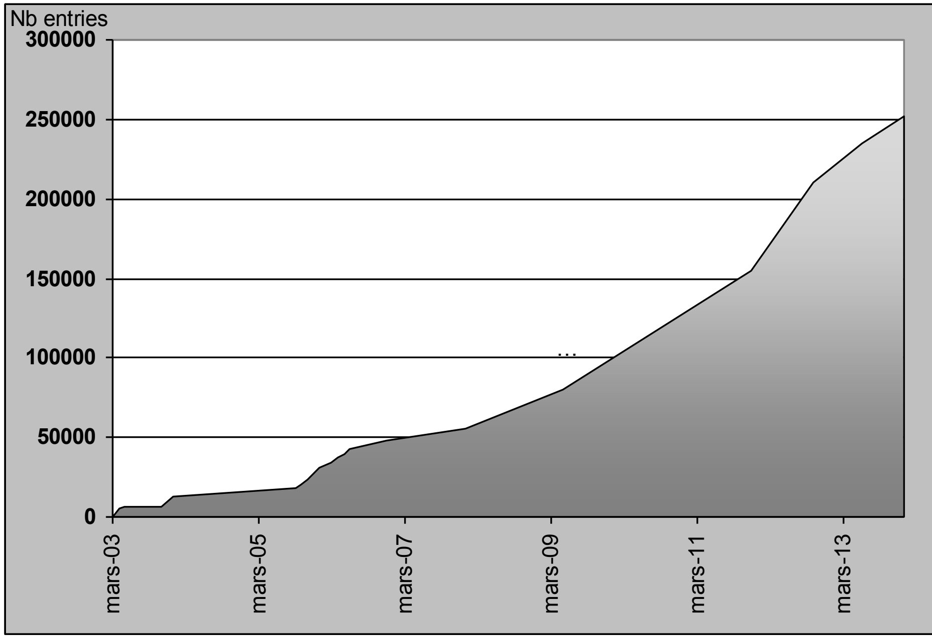
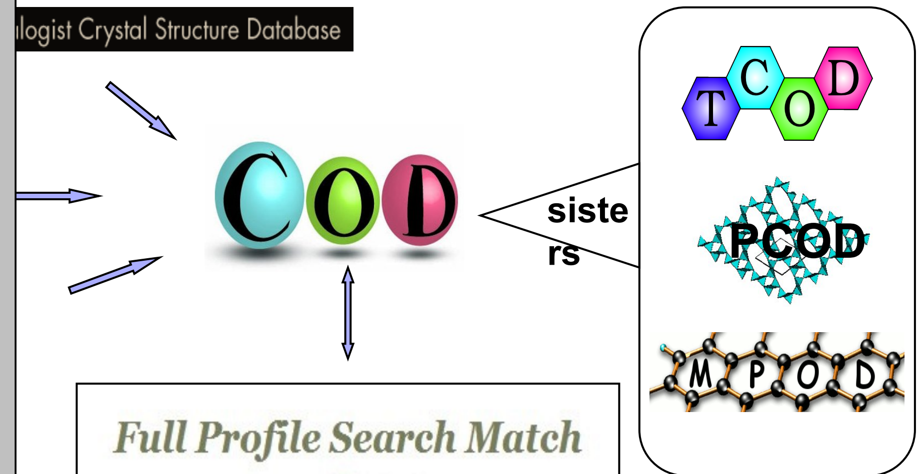


Figure 3: PAIS, preservation process and data lifecycle

Crystallography Open Databases and Preservation: a World-Wide Initiative

Daniel Chateigner (for the COD Advisory Board)





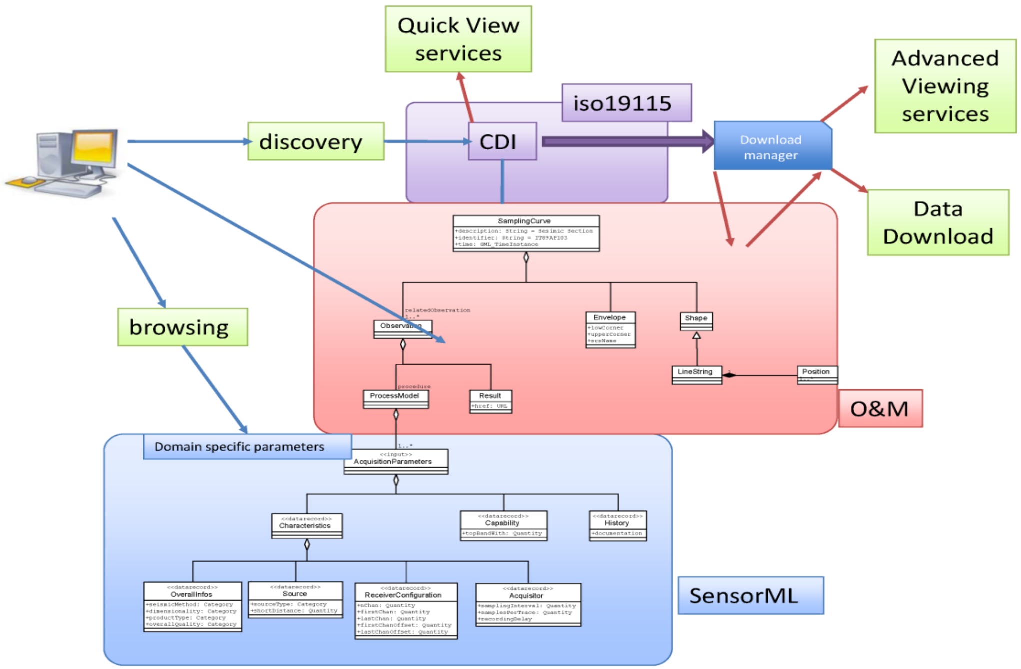
"…there is not yet sufficient coherence of experimental metadata standards or national policy to rely on instrumental facilities to act as permanent archives;
-there is not sufficient funding for existing crystallographic database organisations (which maintain curated archives of processed experimental 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.

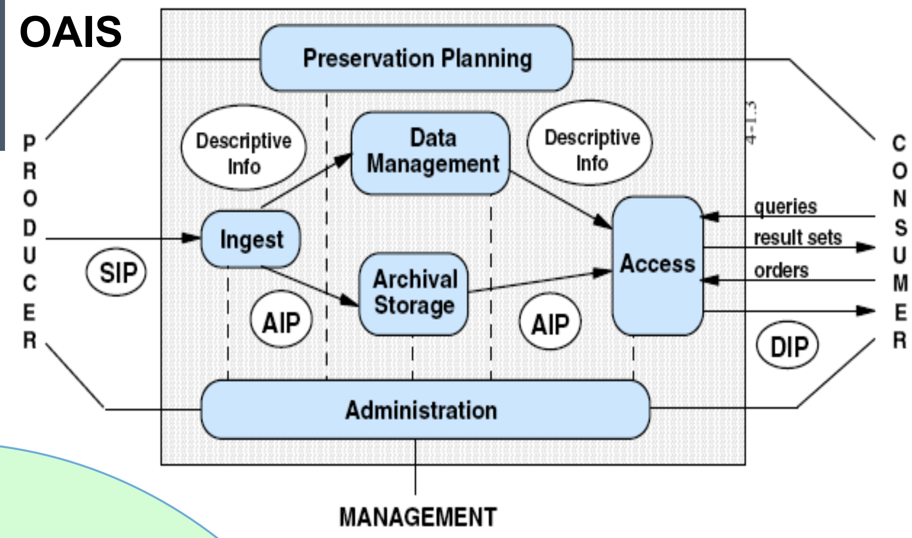


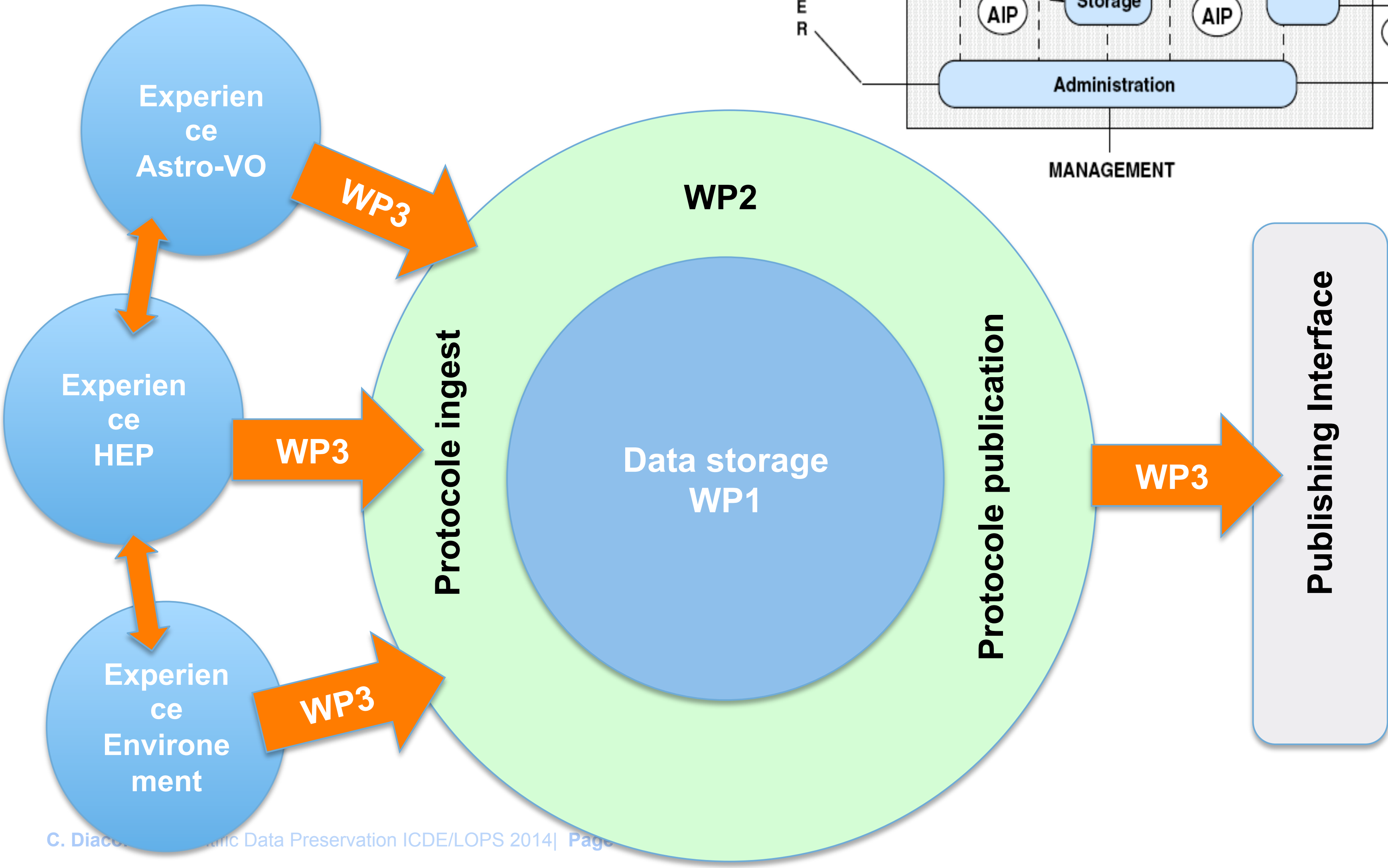
Scientific Data Preservation, Copyright and *Open Science*

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)

Towards a demonstrator



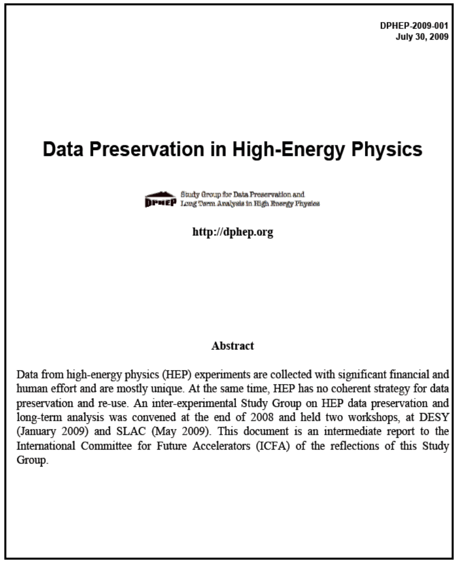


Conclusions

- Research data in the recent period is a « gold mine » with unexploited research potential and therefore shoudl be preserved long term
- The preservation is worth the effort and enables low cost science
- International cooperation is essential
- Frontier technologies and research are involved
- Innovative sollutions can emerge from multidisciplinary initiatives

DPHEP Intermediate Recommendations (end 2009)

- ***arXiv:0912.0255***



DPHEP-2009-001
July 30, 2009

Data Preservation in High-Energy Physics

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

http://dphep.org

Abstract

Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. At the same time, HEP has no coherent strategy for data preservation and re-use. An inter-experimental Study Group on HEP data preservation and long-term analysis was convened at the end of 2008 and held two workshops, at DESY (January 2009) and SLAC (May 2009). This document is an intermediate report to the International Committee for Future Accelerators (ICFA) of the reflections of this Study Group.

- An urgent and vigorous action is needed to ensure data preservation in HEP
  - Many examples for the physics case explored
  - Data is rich and can be further exploited in most cases beyond the collaboration lifetime
- The preservation of the full analysis capability of experiments is recommended, including the preservation of reconstruction and simulation software
- An interface to the experiment know-how should be introduced: **data archivist** position in the computing centres
- The preservation of HEP data requires a synergic action: collaborations, laboratories and funding agencies
- An International Data Preservation Forum is proposed as a reference organisation. The Forum should represent experimental collaborations, laboratories and computing centres

DPHEP publication 2012

> Full status report of the activities of the DPHEP study group, including:

- Tour of data preservation activities in other fields
- An expanded description of the physics case
- Defining and establishing data preservation principles
- Updates from the experiments and joint projects
- FTE estimates for these and future projects
- Next steps to establish fully DPHEP in the field

DPHEP-2012-001
May 2012

Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics

www.dphep.org

Abstract

Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. An inter-experimental Study Group on HEP data preservation and long-term analysis was convened as a panel of the International Committee for Future Accelerators (ICFA). The group was formed by large collider-based experiments and investigated the technical and organisational aspects of the HEP data preservation. An intermediate report was released in November 2009 addressing the general issues of data preservation in HEP. This paper includes and extends the intermediate report. It provides an analysis of the research case for data preservation and a detailed description of the various projects at experiment, laboratory and international levels. In addition, the paper provides a concrete proposal for an international organisation in charge with the data management and policies in high-energy physics.

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

arXiv:1205.4667

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)

EU Visions


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…

Data Preservation in a multidisciplinary context

- **More Coordination**: The organisation should be brought to a long-term perspective by solid, commensurate and courageous decisions of the funding and coordination bodies responsible for the wealth of HEP experimental data produced so far.

- **More Standards** An increased standardisation will increase the overall efficiency of HEP computing systems and it will also be beneficial in securing long-term data preservation.

- **More Technology:** These new techniques (virtualisation etc.) seem to fit well within the context of large scale and long-term data preservation and access.

- **More Experiments**: The expansion of the DPHEP organisation to include more experiments is one of the goals of the next period.

- **More Cooperation: Cooperation with other fields in data management: access, mining, analysis and preservation; appears to be unavoidable and will also dramatically change the management of HEP data in the future.**