

Associer *Invenio* avec  
*Archivematica* pour une  
archive numérique au CERN

JY Le Meur

CERN Digital Memory project leader

Que  
voulons-nous  
préservé ?

Comment  
pouvons-nous  
préservé ?

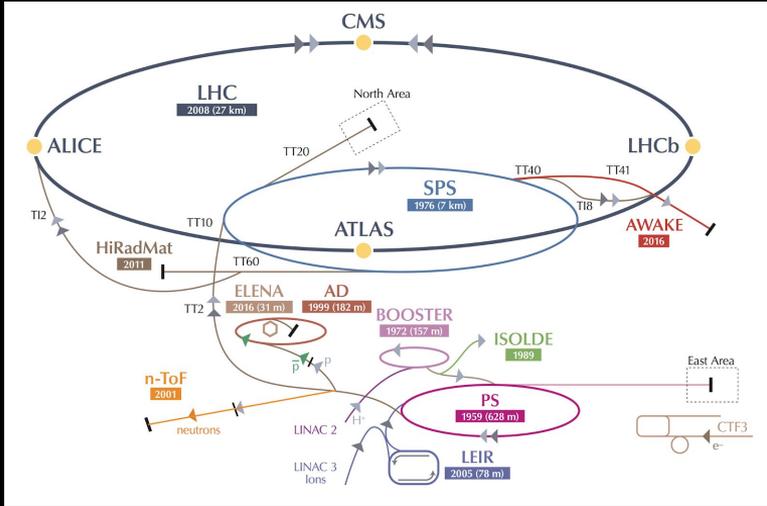
Qu'avons-nous  
développé ?

Vers où  
allons-nous ?

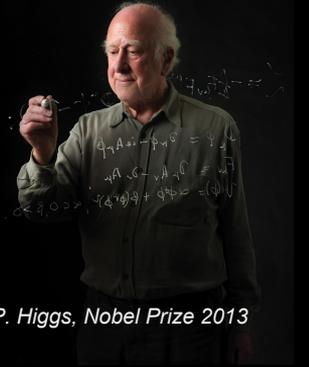


# La physique des particules

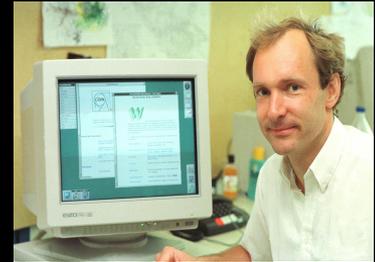
## CERN (1954 -)



LHC: 100 m underground



P. Higgs, Nobel Prize 2013



T. Berners-Lee, Alan Turing Prize 2018



# Les données de la physique



- **Uniques et “chères”**
- **Complexes**, multi-niveaux (4), destinées aux analyses
- **Massives**, par exemple:
  - 100 TB/expérience au LEP
  - 1-10 PB/exp au HERA, TEVATRON ou BaBaR
  - X00 PB/exp au LHC
  - >10 EB à prévoir au HL-LHC
- **Open Access / Data / Science**
  - Exigence de Preservation selon le principe FAIR

**Education**

The CMS Compact Muon Solenoid experiment is one of the large general-purpose detectors built on the Large Hadron Collider (LHC). Its goal is to investigate a wide range of physics, such as the measurement of the Higgs boson, exotic dimensions of dark matter.

Explore CMS >

ALICE (Large Ion Collider Experiment) is a heavy-ion detector designed to study the physics of strongly interacting matter at extreme energy densities, where subatomic matter is produced in dense forms. More than 1000 scientists are part of the collaboration.

Explore ALICE >

The ATLAS (Large Hadron Collider) experiment is a general-purpose detector designed to probe the properties of the Higgs boson, search for new particles, and study the production of supersymmetric particles and other phenomena at the LHC.

Explore ATLAS >

The LHCb Large Hadron Collider beauty experiment aims to record the decay of particles containing b and c quarks, known as B mesons. The detector is designed to gather information about the beauty, charm, top quark and anti-top quark.

Explore LHCb >

For education purposes, the content of many data sets to be processed in a format designed to be easy to use for simple applications. Get in touch if you wish to build your own applications similar to those shown here.

Visualizations >

Learning Resources >

## CERN Open Data & Analysis Preservation

## Le projet DP-HEP

Garantie 30 ans?

CERN Accelerating science

Sign In Directory

**DP-HEP** Data Preservation in High Energy Physics

Collaboration for Data Preservation and Long Term Analysis in High Energy Physics

Partners Accelerators Meetings ICFA Study Group About Us

FOLLOW THE LINKS BELOW TO FIND INFORMATION ON OUR PARTNER ORGANIZATIONS. EACH REPRESENTS SOME EXPERIMENTS AND ACCELERATORS TO THE COLLABORATION FOR DATA PRESERVATION IN HIGH ENERGY PHYSICS.

Search this site Search



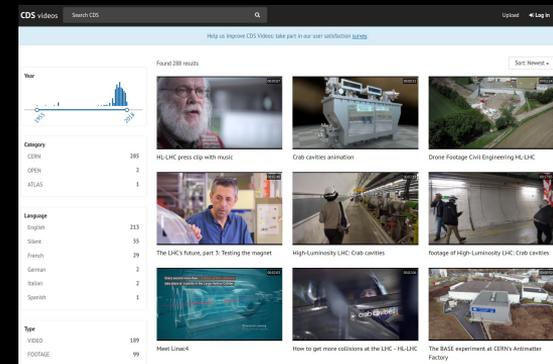
# Les données du patrimoine



- Publications et preprints
- 1'740 Kg de bandes audio
- 6'000 video tapes & films
- 450'000 images: négatifs, formats medium & large, diapos



Le Serveur de Documents  
du CERN



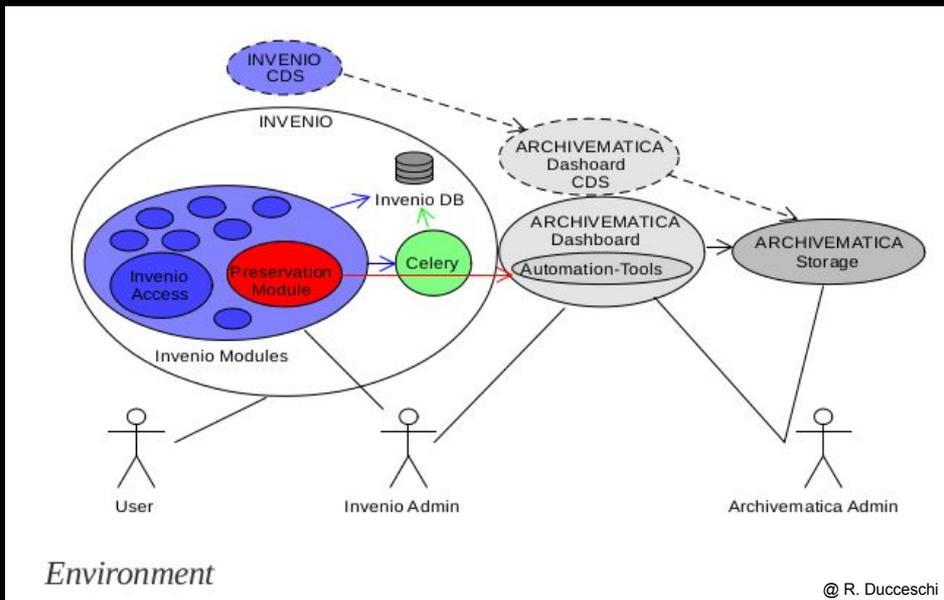
Comment  
pouvons-nous  
préserver ?



# Devenir un “trustworthy digital repository” (TDR)

- **Le centre de données du CERN:**
  - Focalisé sur le “*bit preservation*”
- **Une tentative de certification ISO 16363**
  - Encouragée en 2013 par le Comité Européen de Stratégie pour la Physique des Particules
- **Les métriques** sont complétées, excepté:
  - ‘business continuity’ et ‘disaster preparedness’
  - Mise à jour de la ‘CERN-wide policy’
  - **La création de AIPs** (Archival Information Packages) conforme au modèle OAIS

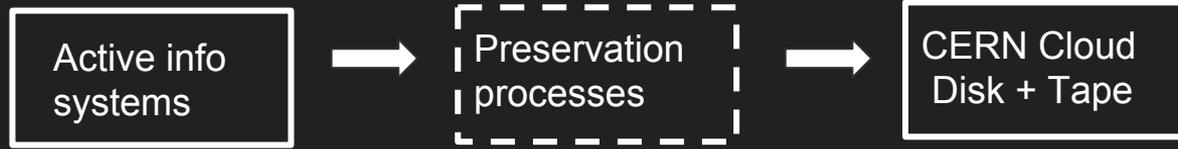
# L'environnement



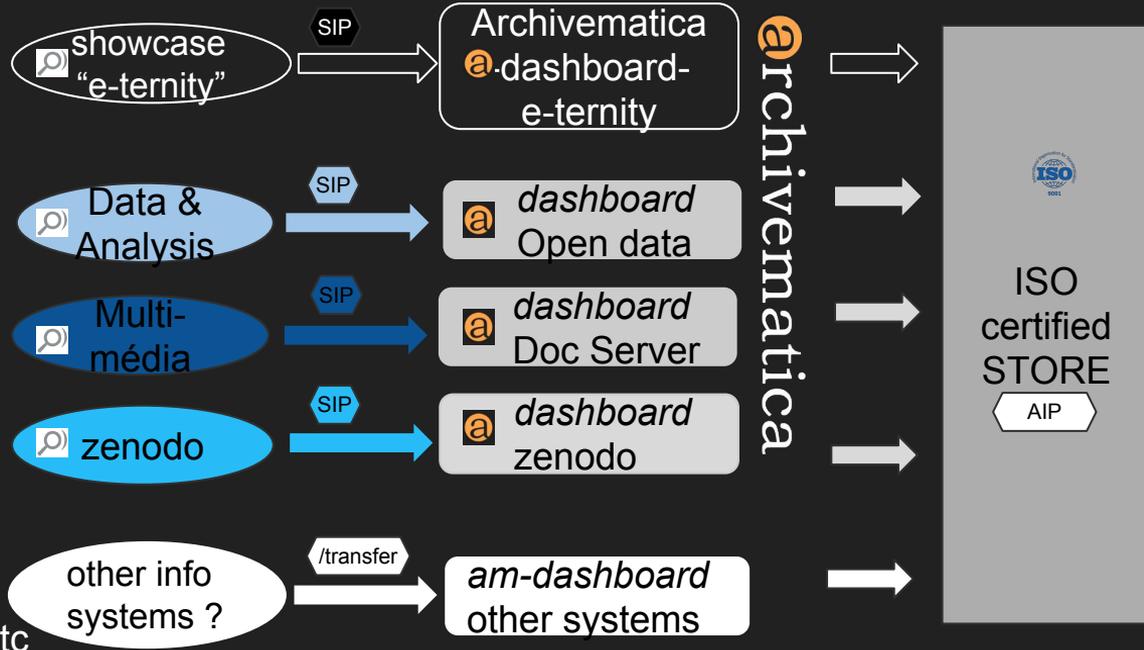
- Les logiciels de dépôts numériques
  - Invenio: [invenio-software.org](https://invenio-software.org)
- Les logiciels de préservation
  - Archivemata: [archivemata.org](https://archivemata.org)
- Au CERN: des systèmes d'information éparpillés



# Les flux



INVENIO



- Twiki
- Drupal
- Indico etc

Qu'avons nous  
développé ?



# Archivematica dans le Cloud CERN

- L'alignement sur l'infrastructure ✓
  - Développement et testing en local avec Docker
  - Déploiement de VM sous Puppet : AM v1.6  
[github.com/CERN-E-Ternity/archivematica-puppet](https://github.com/CERN-E-Ternity/archivematica-puppet)
  - La gestion des versions (et des conflits, v1.8 en cours)
- La gestion des permissions et les transferts de fichier ✓
  - Shared filesystem VS scp VS xrootd
- L'appel aux services centralisés
  - Remplacement de SQLite par MySQL (DB on demand)
  - ElasticSearch on Demand
  - OAuth





# Des SIPs aux AIPs

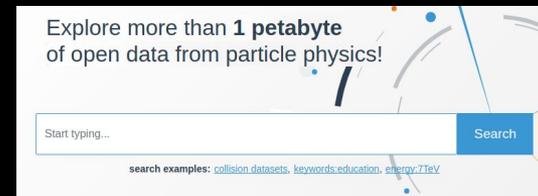


- Invenio-SIPStore: création des ‘Submission Information Packages’
  - Selon les specs du [BagIt File Packaging Format](#) (IETF)
  - Avec un délai selon les instances
  - Métadonnées partielles en Dublin Core
    - en PREMIS/METS dans le AIP
    - et un fichier complet de métadonnées (JSON)
  - Module: [github.com/inveniosoftware/invenio-sipstore](https://github.com/inveniosoftware/invenio-sipstore)
- Invenio-Archivematica: automation du transfert avec les APIs
  - Conversion de formats selon les dashboards
  - Échange du ‘statut’ du processus d’archivage
  - Module: [github.com/inveniosoftware/invenio-archivematica](https://github.com/inveniosoftware/invenio-archivematica)





# Des extensions nécessaires



- Le défi des fichiers révisés
  - Analysis Preservation: des petits fichiers très nombreux >5000
  - Open Data: ~600'000 fichiers de plusieurs GB chacun attachés à ~6000 records

→ Re-ingestion d'une notice dans le cas d'un seul fichier modifié: nécessité de fichiers "**externes**" dans la structure du SIP Store (fetch.txt)
- Des fichiers gigantesques
  - Le passage dans le cache d'Archivematica (x 3)
  - Le temps de transfert d'un filesystem à l'autre

→ Ajout de la notion de "**Weak Archiving**":

  - Fichiers non inclus dans l'AIP mais traités pour 'bit preservation'

Où allons-nous ?



# Conclusions

---

- Vers une archive OAIS pour les données scientifiques et patrimoniales
- Depuis les systèmes d'information sous Invenio vers une 'Dark Archive'
  - sans création systématique de DIP
  - avec une interférence minimale des deux systèmes
- Deux modules spécifiques: Invenio-SIPStore et Invenio-Archivematica
- Des défis de "taille"
  
- D'autres initiatives en cours:
  - Projets EU: ATTRACT et ARCHIVER
  - Au CERN prochainement:
    - Archivematica Camp **22-24 Oct 2019**
    - PV2020 Conference **12-14 May 2020**
  
- CERN est devenu membre de la Data Preservation Coalition