FOSTER

Facilitate Open Science Training for European Research

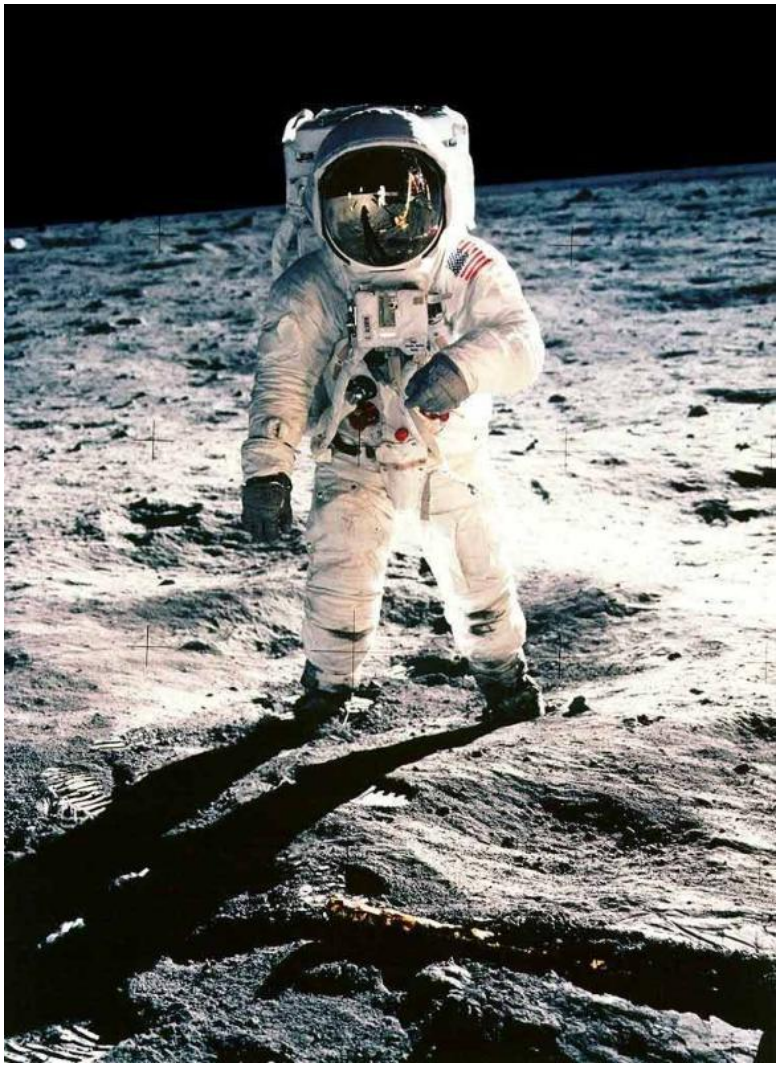
La conservation pérenne des données de recherche

Lorène BECHARD

Centre Informatique National de l'Enseignement Supérieur

30 juin 2015

Un exemple…

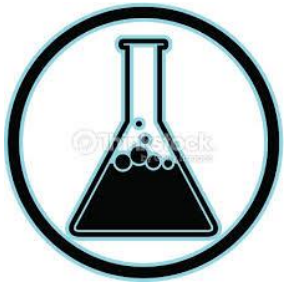


FOSTER

Sommaire

- Définitions et contexte juridique

- Les données scientifiques, quelles spécificités ?



- L'archivage électronique, késaco ?

- Un acteur ESR désigné, le CINES

FOSTER

Donnée VS Archive

- Une donnée / un document est :
  - produit dans le cadre d'une activité,
  - sous une forme bien spécifique,
  - à une date donnée et dans un lieu précis,
  - sur un support désigné.

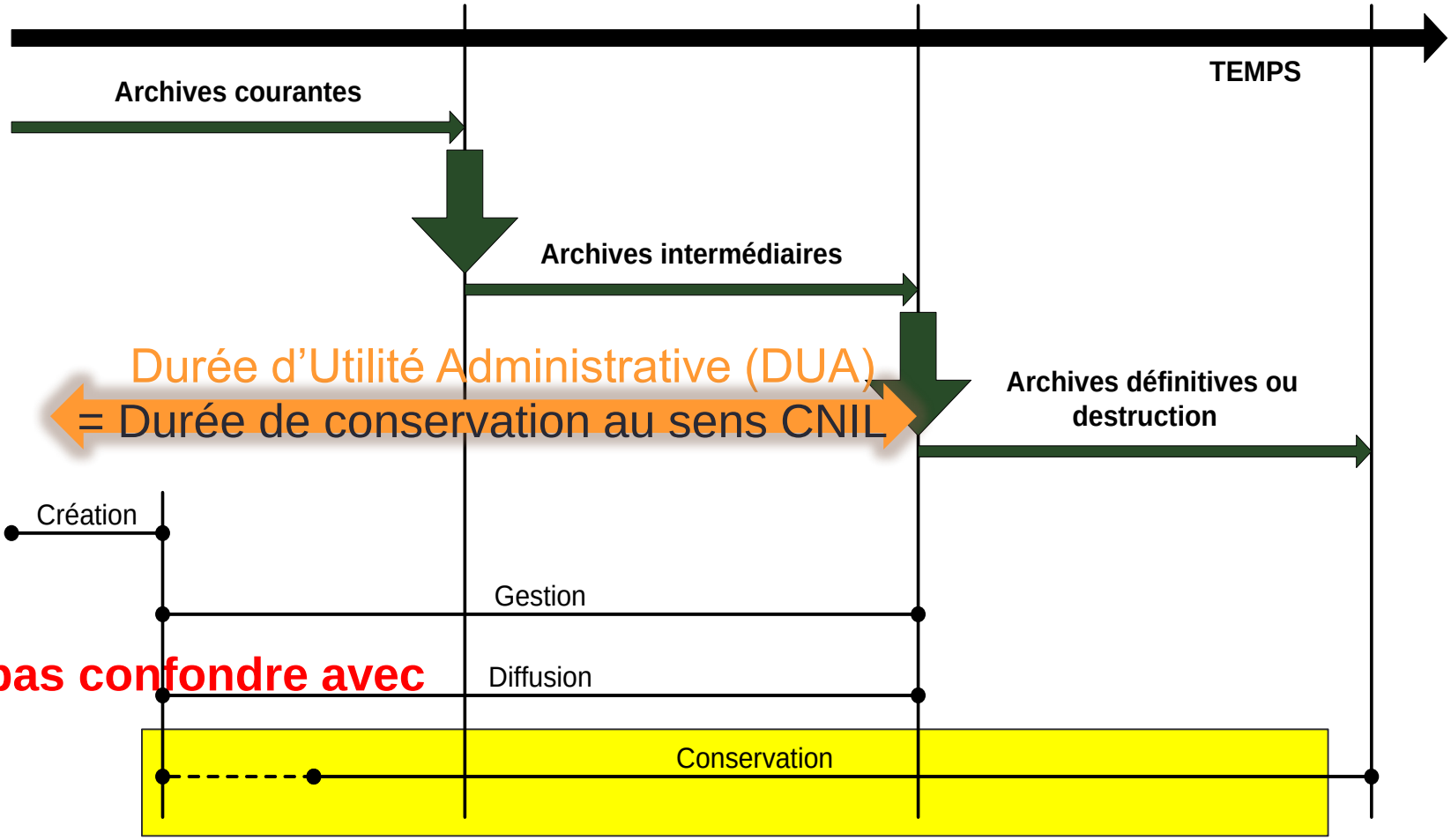
→ C'est donc une « archive en devenir ».

« Les archives sont l'ensemble des documents, quels que soient leur date, leur lieu de conservation, leur forme, leur support, produits ou reçus par toute personne physique ou morale, et par tout service ou organisme public ou privé dans l'exercice de leur activité. »

Art.L 211-1 du Code du Patrimoine

FOSTER

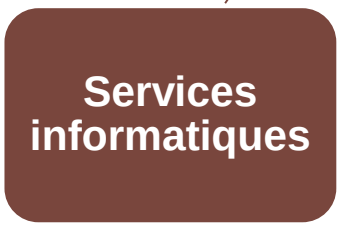
Le cycle de vie de la donnée



À ne pas confondre avec

FOSTER

L'archivage dans H2020 : les acteurs

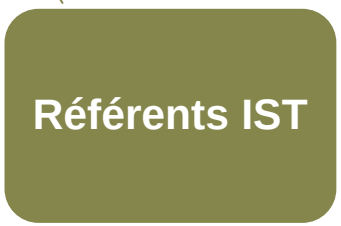


Expertise sur les méthodes et moyens sécurisés de stockage et d'accès aux données numériques

Connaissance du contexte de production, expertise sur le contenu informationnel des données et leur usage, description des données

Pilotage du DMP, identification de la propriété des données, allocation des ressources nécessaires

Expertise du cycle de vie des données, leur évaluation, sélection et organisation pour l'archivage



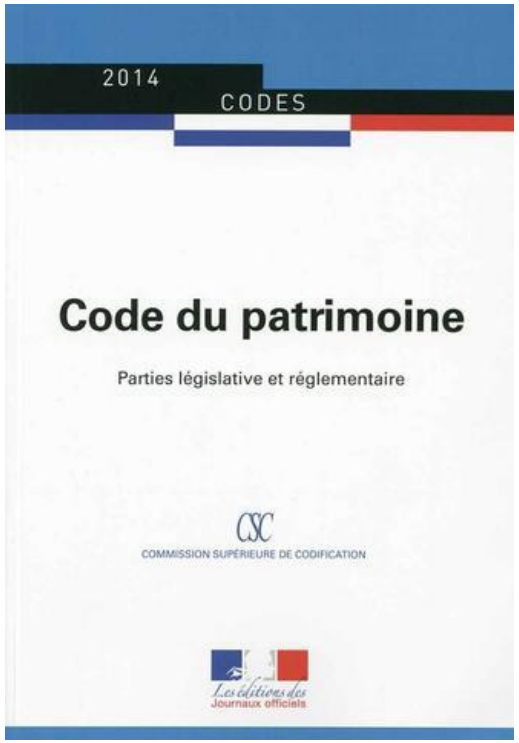
Expertise sur le partage, la diffusion et la réutilisation des données

FOSTER

L'archivage dans H2020 : quelles responsabilités ?

- Identifier la propriété des données produites
  - Déterminée dans l'accord de consortium ?
  - Porteur de projet français = données relevant de la législation sur les archives publiques ?

- Une archive est publique le plus souvent, car produite :
  - par des organismes publics,
  - et/ou dans le cadre d'une mission de service public



Une donnée publique n'est pas forcément publique !

Obligations légales pour les archives intermédiaires

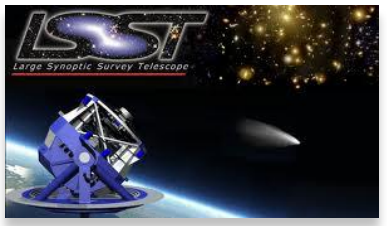
- Gestion à la charge des producteurs / administrations productrices
  - Service dédié pour la gestion des archives
  - Doté de moyens suffisants
  - Externalisation possible mais encadrée
- Quelles responsabilités ?
  - Traiter tous les documents produits, y compris les archives électroniques
  - Suivre les prescriptions en termes de classement et de conservation
  - Respect des règles de communicabilité établies par la loi
- Tri à l'issue de leur durée d'utilité administrative (DUA)
  - Collaboration entre producteurs et archivistes
- Archives à éliminer → visa obligatoire de l'administration des archives
- Archives définitives → conservation dans les services d'archives publics

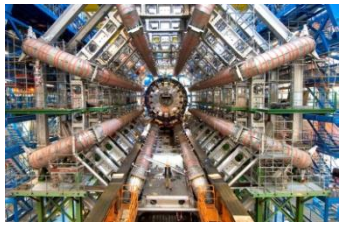
FOSTER

→ Contrôle scientifique et technique effectué notamment par le service interministériel des archives de France (SIAF)

Panorama de la donnée scientifique numérique

- Démocratisation de la donnée (openData)
- Explosion du volume des données :
  - Nouveaux capteurs (plus précis)
    - LSST : Large Synoptic Survey Telescope (15 à 30 Térabytes par nuit).
    - LHC : Le Grand collisionneur de hadron (15 petabytes par an)
  - Augmentation des capacités de calcul
  - Champs de recherche de plus en plus larges
- Exploitation
  - Interdisciplinarité : interdépendance des thématiques scientifiques
  - Data Mining : recherche d'information cachée
  - Outils de visualisation et Web 2.0





La donnée scientifique est rapidement confrontée aux problématiques du BIG DATA

FOSTER

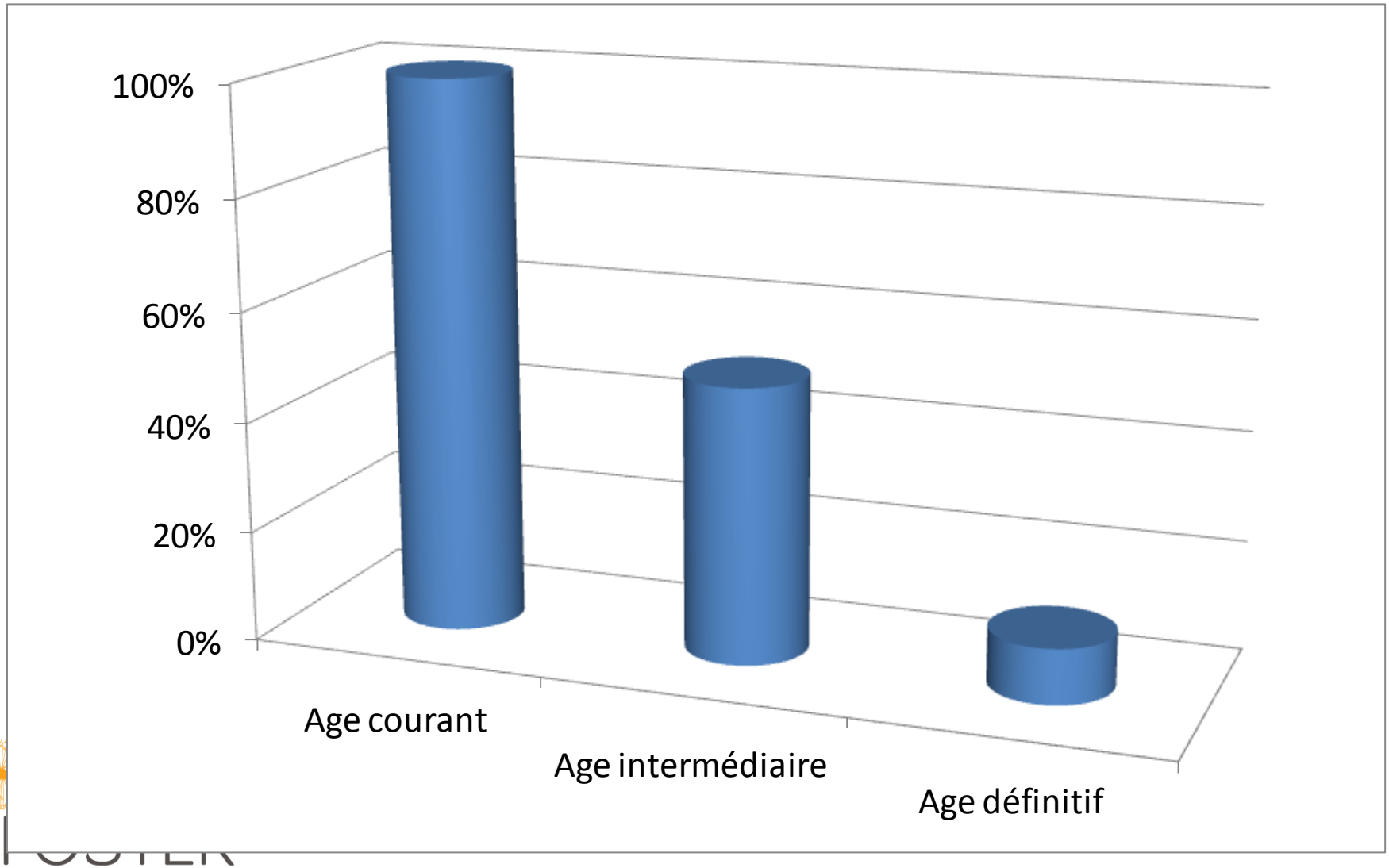
Panorama de la donnée scientifique numérique

… mais avec des préoccupations supplémentaires dans une perspective d'archivage

- Formats de fichiers complexes et variés
  - Quelques formats « pivots »
    - HDF
    - NetCDF
  - Beaucoup de formats « maison » binaires
- Absence de documentation autour des données
  - Indispensable à la compréhension pour une utilisation future
  - Nécessaire collaboration producteurs / archivistes

FOSTER

Quantité de données scientifiques conservées



Pourquoi conserver une archive ?

- Pour des raisons
  - **administratives :** pour faciliter le travail quotidien des agents et des usagers (le diplôme d'un étudiant, le dossier de carrière d'un personnel…)
  - **juridiques :** pour justifier d'une action ou d'une activité lorsqu'il y a contentieux (contentieux en édition, plagiats…)

…pendant la période d'archivage intermédiaire (DUA)

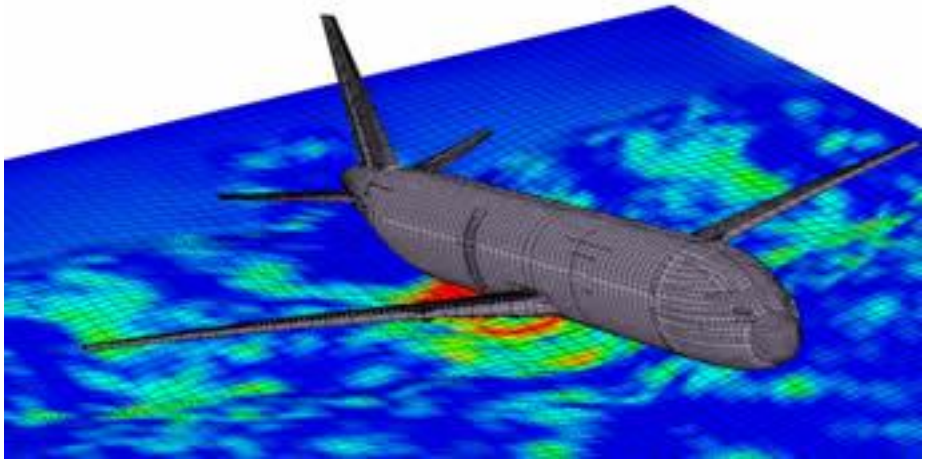
- Pour des raisons historiques (pour témoigner de l'activité d'un organisme, d'une personne, d'une équipe) pour un archivage définitif

FOSTER

Pourquoi conserver une donnée scientifique ?

- Pour la réutiliser
  - réaliser des statistiques, reprise de calculs, traduction des résultats en image (lisibilité, fiabilité)
  - Prouver les résultats scientifiques obtenus (intégrité authenticité)

- Pour la protéger (open data vs confidentialité)



FOSTER

La problématique de l'archivage numérique

Qu'est-ce que l'archivage électronique ?

L'archivage des documents électroniques consiste à conserver le document et l'information qu'il contient :

- *dans son **aspect physique** comme dans son **aspect intellectuel**,*
- *aussi longtemps que nécessaire (**moyen et long termes**),*
- *de manière à ce qu'il soit en permanence **accessible et compréhensible**.*

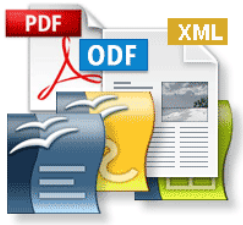
FOSTER

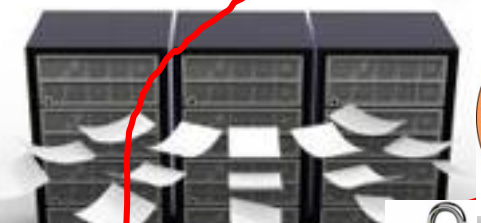
Sauvegarde VS Sauvegarde sécurisée VS Archivage

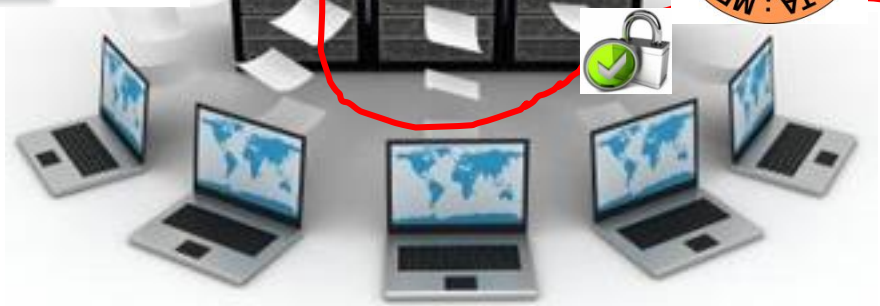
ARCHIVAGE

SAUVEGARDE

SAUVEGARDE SECURISEE







- ✓ Respect de l'intégrité et de l'authenticité
- ✓ Garantie de lisibilité
- ✓ Garantie de compréhension

→ Sur le moyen et long terme

FOSTER

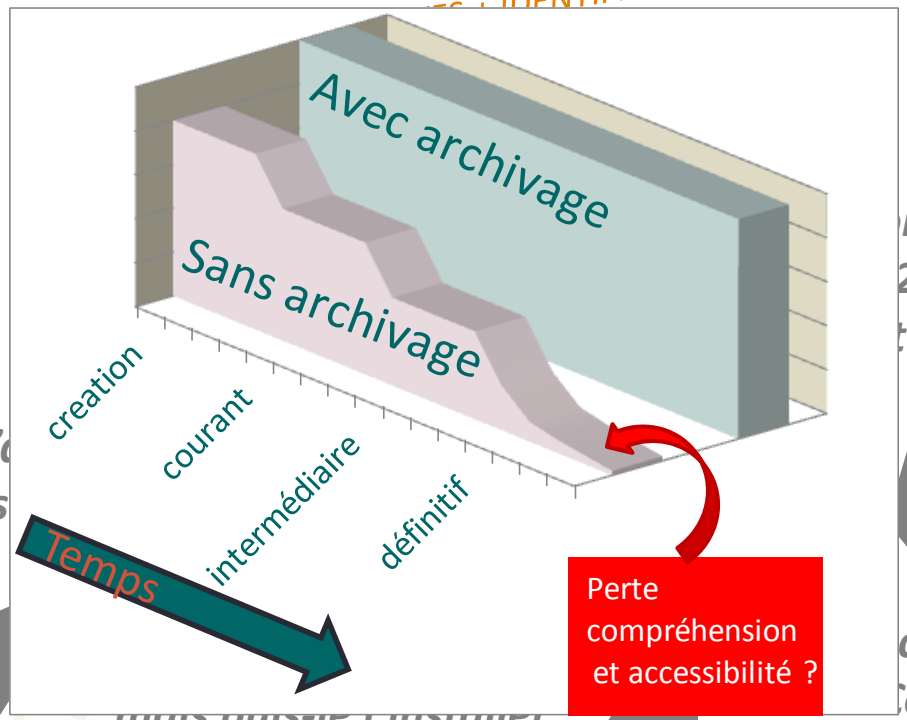
Voici un document que j'ai créé en 1998…

De quoi s'agit-il déjà ? Est-ce bien ce qui est indiqué sur la disquette ?

IDENTIFICATION

La disquette est-elle toujours en bon état ?

SUPPORT : VEILLE + MIGRATION PHYSIQUE



…n portable, acheté …2013, n'a pas de …teur de disquette…

ENVIRONNEMENT MATERIEL : VEILLE TECHNO et ANTICIPATION

Ça marche ! Mais j'… perdu toute ma mis… en forme…

INTEGRITE AUTHENTICITE

…ce document avec Claris …Comment retrouver ce logiciel ? Quel est le format du document ?

ENVIRONNEMENT LOGICIEL : privilégier les FORMATS DURABLES + MIGRATION LOGIQUE

…mais puis-je l'installer et l'utiliser sous Windows 8 ?

SYSTÈME D'EXPLOITATION

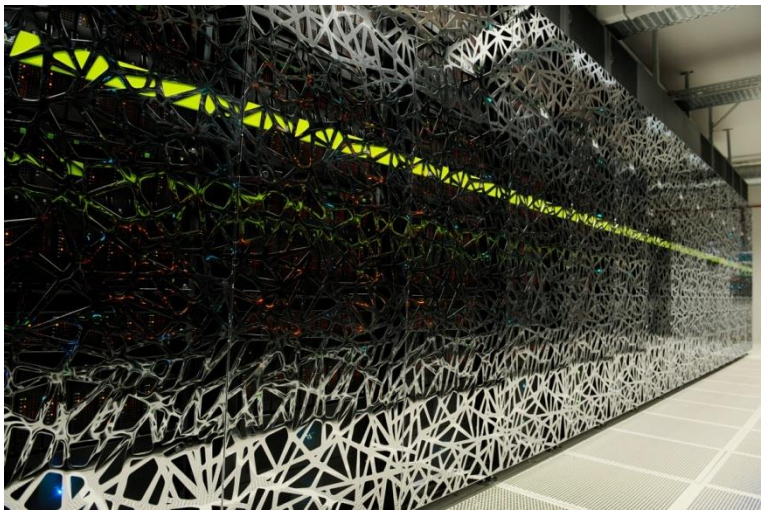
FOSTER

Services de « préservation à moyen et long termes » de données scientifiques du CINES pour une **prise en charge** de la problématique **tout au long du cycle de vie du document**

Le CINES et sa mission de préservation numérique

Centre Informatique National de l'Enseignement Supérieur

- Basé à Montpellier (Hérault, France)
- EPA créé en 1999, succédant au CNUSC (Centre National Universitaire Sud de Calcul) – créé en 1980
- Placé sous la tutelle du Ministère chargé de l'Enseignement Supérieur et de la Recherche
- Missions
  - Calcul numérique intensif
  - Archivage pérenne de données électroniques
  - Hébergement de matériels informatiques à vocation nationale
- Plus d'informations : http://www.cines.fr/



La machine " Occigen" du CINES (2,1 Pflops – 26ème du Top 500)

L'archivage au CINES en quelques dates…

2004

- 1ères **réflexions** sur l'archivage numérique
- **Prestation de conseil** en AE (2005)

2006

- **Arrêté du 7 août 2006 (Thèses)**
- 1ères **thèses versées** dans PAC_V1, **une plateforme d'archivage « maison »** (arrêté du 07/08/2006)
- Convention pour l'**archivage du portail Persée** (02/05/2007)

2008

- **Recadrage des activités du CINES** (lettre du 12/02/2008)
- **Mise en production de PAC_V2** (basée sur la solution Arcsys d'Infotel et un client spécifique)
  → *« migration de plateforme »*

FOSTER

L'archivage au CINES en quelques dates…

2010

- **Archivage de HAL** (Hyper Articles en Ligne), des **données** orales du **CRDO** (TGE-Adonis), de **livres numérisés (Cujas, BIUS, BSG…)**
- 2009 : **1er To archivé**
- **Agrément SIAF** (14/12/2010)
- Projet APARSEN : **évaluation norme ISO** 16363 (2011)

2012

- Publication de **guides méthodologiques sur les formats de fichiers** audio et vidéo (2010-2012) + Bases de données (2013) + PDF (en cours)
- **Archivage** des données **de la Cour des Comptes conformément au SEDA** (2013)
- **Archivage** des données **de l'INSERM** (2013)

2014

- Fin 2013 / début 2014 : **Archivage de données scientifiques dans le nœud EUDAT@CINES (projet européen FP7)**
- Début 2014 : **Renouvellement agrément SIAF**
- 1er janvier 2014 : **25To** utiles archivés dans PAC
- **Modification des statuts du CINES** : mission d'archivage pérenne inscrite

FOSTER

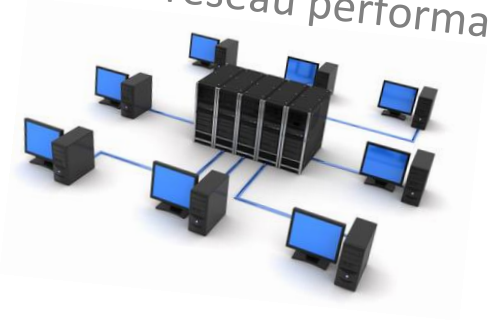
Des équipements exceptionnels

Des bâtiments dédiés :

- 5 salles machines : 1 400 m²
- Locaux techniques : 2 000 m²

Des équipements informatiques :

- Un supercalculateur de niveau mondial
- Capacités de stockage de plusieurs PetaOctets
- Des accès réseau performants



Des équipes d'experts :

- Système
- Réseaux
- Sécurité
- Bases de données
- Stockage
- Calcul haute performance, visualisation
- Préservation (archivistes, réseau de professionnels de l'IST…)
- Administrateurs

FOSTER

Des solutions techniques et organisationnelles

Nos solutions sont :

- Personnalisées
- Mutualisées
- Economiques
- Sécurisées
- Performantes
- Standardisées
- Complémentaires de vos solutions de diffusion (Diffusion vs. Archivage)

→ Seul un centre national spécialisé peut proposer cette qualité de service

FOSTER

Les infrastructures d'archivage du CINES

PAC

Archivage à long terme de données scientifiques, patrimoniales, administratives

Assurance qualité
OAIS (ISO 14721)
Expertise format
Compétences archivistiques
Processus métier
Gestion des risques

ISAAC

Archivage intermédiaire de données scientifiques

EUDAT

Archivage de données scientifiques pour des communautés européennes structurées

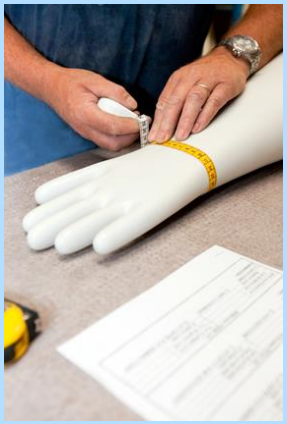
FOSTER

Le processus d'archivage au CINES

Plateforme d'archivage (agréée SIAF + Santé + DSA + ISO 16363)

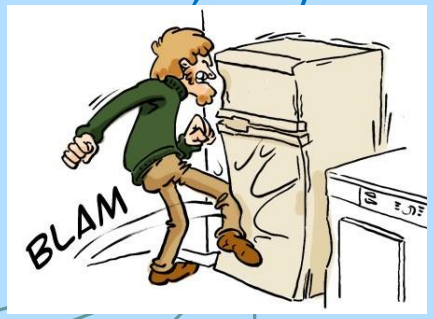
1. Réception

2. Vérification de la qualité des données reçues

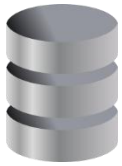
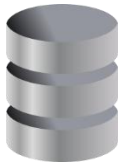


3. Ajout d'informations (PID, empreintes, date d'archivage…)

4. Traitements complémentaires (récupération d'informations associées, etc.)



5. Stockage de l'archive en plusieurs exemplaires

Salle machine 1	Salle machine 2	Salle machine 3	Site distant (> 300km)
			

6. Vérification périodique de tous les exemplaires archivés

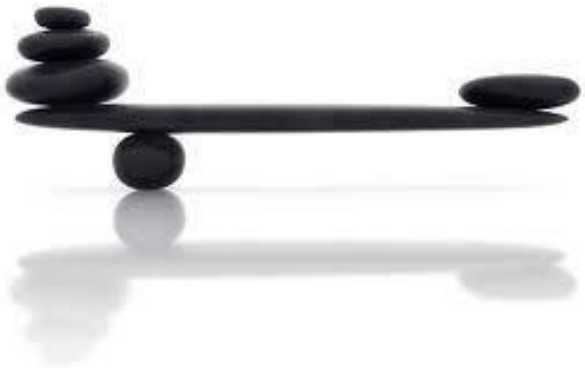
+ **migration** quand il y a lieu

Un projet d'archivage au CINES, c'est… ?

- **L'archivage de données produites par les organismes de l'ESR :**
  - ✓ ***validées*** *(archivage courant, intermédiaire et définitif)*
  - ✓ ***sélectionnées***
  - ✓ ***documentées*** *(jeu minimal de métadonnées : Dublin Core / SEDA)*

Quel travail de curation sur les données pour l'archivage ?

- *adapté à la durée de conservation des données*
- *et au niveau de services demandé : intégrité, lisibilité, intelligibilité*



FOSTER

Un projet d'archivage au CINES, c'est… ?

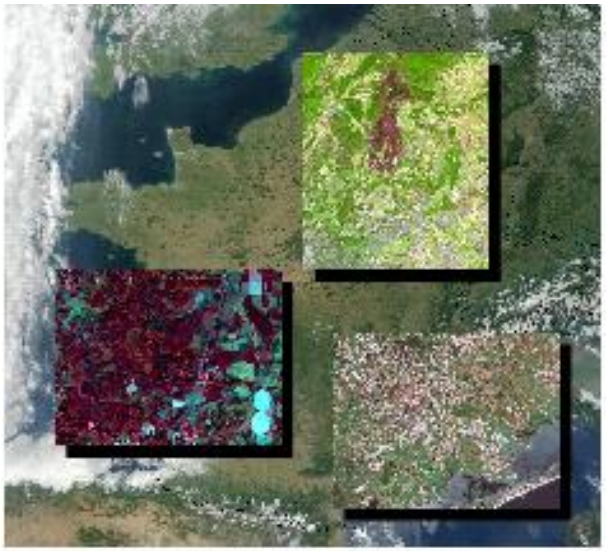
- **Une équipe-projet dédiée :**
  - ✓ *Un référent-projet informatique et un archiviste côté CINES*
  - ✓ *Un référent-projet côté Service Versant*
  - ✓ *Des développements informatiques à prévoir : interfaçage avec la plateforme*

- **Un partenariat encadré :**
  - ✓ *Lettre d'intention*
  - ✓ *Convention d'archivage*
  - ✓ *Tarification au Téraoctet utile archivé et en fonction du niveau de service*

FOSTER

Exemple : « l'archivage des données satellitaires Géosud »

- **Type de données concernées :**
  - Images satellitaires brutes et orthorectifiées
  - Couverture annuelle nationale et régionale, acquisitions en pied d'antenne (SPOT6/7)
  - Eléments de volumétrie : environ 10 To/an
- **Niveau de service assuré :**
  - Archivage pérenne avec conservation sur le long terme
  - Différents niveaux de métadonnées de description : Dublin Core, ISO 19115
  - Contrôle de la validité des formats des fichiers : GeoTIFF, JPEG2000, XML, PDF
- **Intérêt de l'archivage :** pouvoir [illegible] évolutions des territoires



Problème de conformité à la norme GeoTIFF (système de projection Lambert 93 utilisé pour les images orthorectifiées absent de la norme).
Solution retenue : archivage des images brutes en plus

Eudat « European Data for e-science »

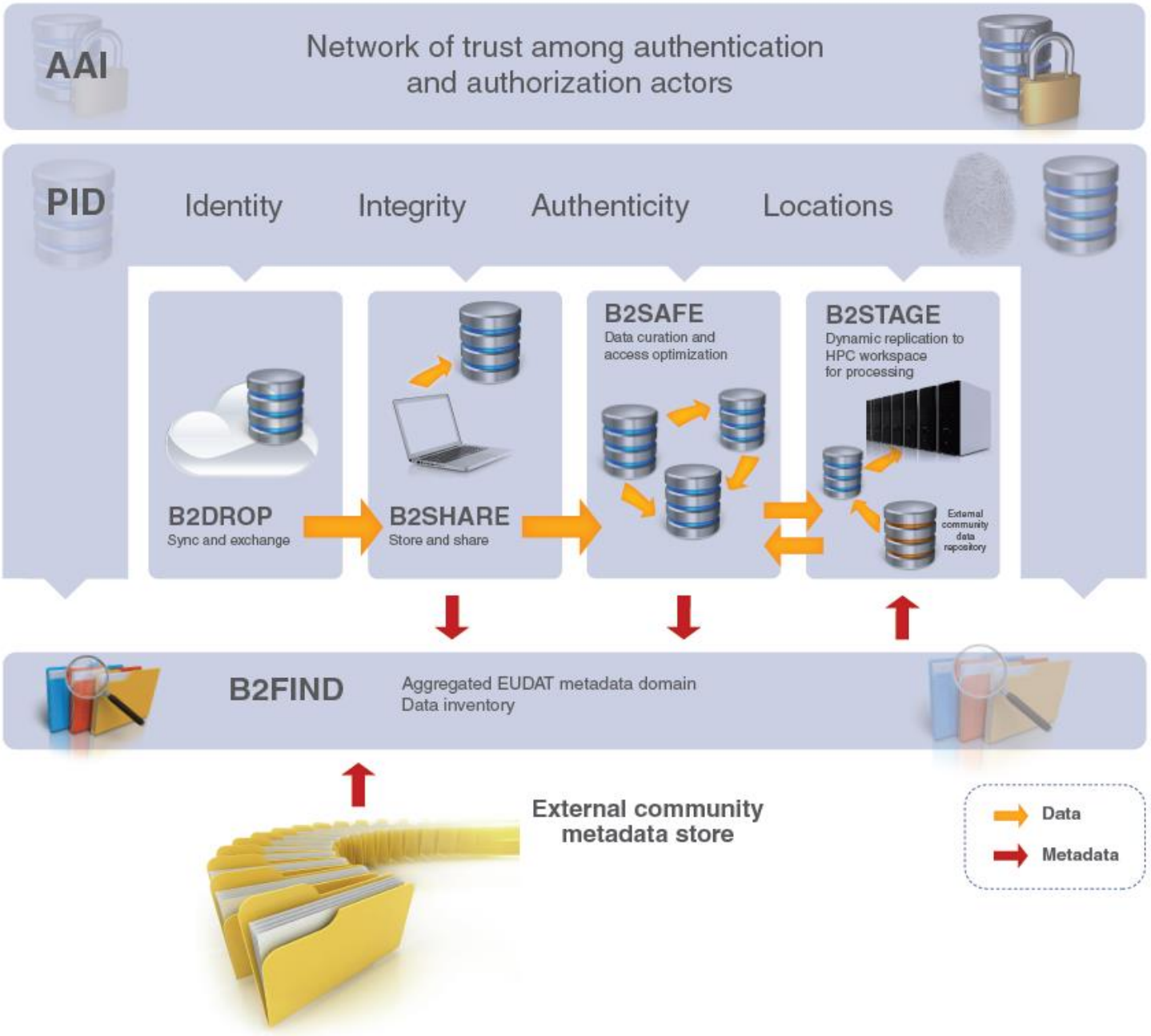
Objectif : fournir une Infrastructure Collaborative de Données (CDI) européenne qui adresse le cycle de vie de la donnée (Stockage, Traitement, Accès, Echange, Conservation à moyen et long termes)

Public cible : Les communautés de la recherche dans toutes les disciplines

Réseau de 35 partenaires (centres de calcul, centres de données, communautés scientifiques) à travers 15 pays européens

→ dont le CINES seul centre de données et de calcul en France

EUDAT



FOSTER

L'appel à collaborations EUDAT

POURQUOI FAIRE ?

- *Data synchronisations and exchange*
- *Data repository and data sharing*
- *Data replication and preservation*
- *Data staging for analysis and processing*
- *Data discovery and search*
- *Data typing and visualization*
- *New services or tools in the area of Big Data Analytics, Semantic web, etc.*

→ **Possibilités de financement par la CE**

POUR QUI ? Toutes les initiatives, infrastructures et communautés de recherche européennes

CRITERES d'EVALUATION :

- Faisabilité technique du pilote (considérant le calendrier et les ressources dédiés) [40%]
- Participation et bénéfices attendus pour la communauté de recherche ciblée [20%]
- Valeur ajoutée pour EUDAT (développement de services, de communautés) [20%]
- Contribution à l'open access [10%]
- Développement de solutions selon des approches génériques telles que RDA [10%]

CALENDRIER :

- Date limite des soumissions : 30/09 – 17h CET
- Implémentation du pilote : 01/01/2016 – 30/06/2017

FOSTER

Les données archivées au CINES

Des questions ?

Plus d'informations : **http://www.cines.fr/archivage/**

lorene.bechard@cines.fr