L’offre de service archivage du CINES

Journée archivage Grenoble

GRICAD – 20/11/2019

Olivier Rouchon

Le Centre Informatique National de l'Enseignement Supérieur

Fournit à la communauté ESR des ressources, services et expertises informatiques exceptionnelles

<u>3 activités principales:</u>

- ✓ Calcul intensif
- ✓ Archivage pérenne
- ✓ Hébergement

✓Basé à Montpellier, France – approx. 60 personnes (ingénieurs, techniciens, administratif)

✓Créé en 1999, initialement **CNUSC** (Centre National Universitaire Sud de Calcul) – créé en 1980

✓Sous la tutelle directe du Ministère de l'Enseignement Supérieur et de la Recherche (MESR)

L’organisation datacentrique

Les données: point central de l’offre de services du CINES

Stockage
Objets
Réplications

Echantillons
pour les
communautés
de recherche

Traitements
Calculs
IA

Archivage
pérenne
Certifications

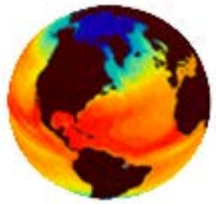
Portails de
diffusion
Open Science

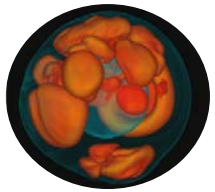
Entrepôts
de données

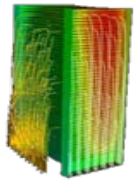
Plan de gestion
des données

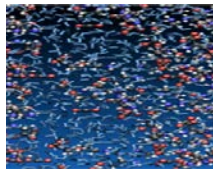
Les données scientifiques

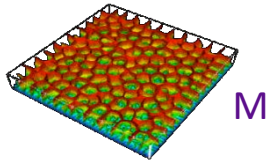
Pour la science

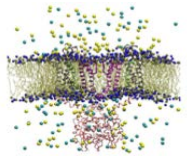

Climat

Astrophysique


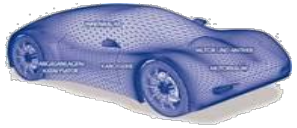

Energie

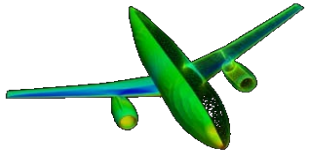
Chimie


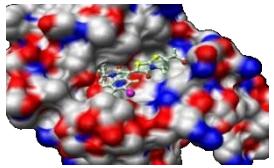

Matériaux

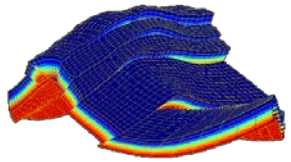
Sciences
du vivant


Pour l'innovation


Automobile

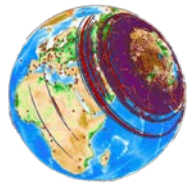
Aéronautique


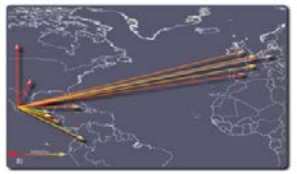

Pharmacologie

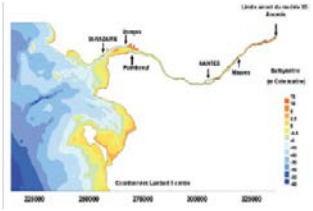
Exploration
pétrolière


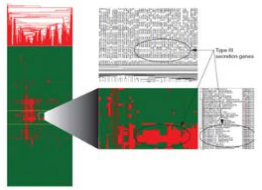

Médecine
personnalisée

Pour l'aide à la décision


Risques naturels

Risques biologiques
et épidémiologiques



Impact des
activités
industrielles

Sécurité


(P. Lavocat, 2016 ; GENCI) – Journée Grands Défis Septembre 2016

Les données archivées

Thèses de doctorat françaises

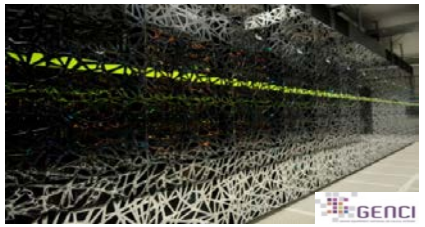
Publications scientifiques

Documents patrimoniaux numérisés

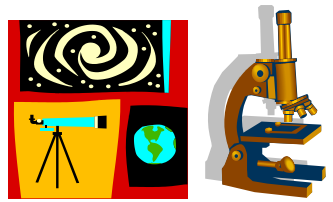
Vidéothèques et banques d'images

Rapports

Données de gestion



Résultats de simulation



Observations

Les données archivées

Données d'observation

•Temps réel,

•Uniques, impossibles à reproduire

Relevés météo, images satellite, enquêtes sociales, fouilles archéologiques

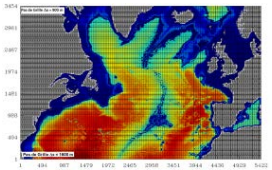
Données expérimentales

•Equipements de laboratoires,

•Reproductibles, parfois coûteuses ou dangereuses

Séquences peptides, poids biomasse

Données de simulation numérique



•Modèles informatiques ou statistiques,

•Reproductibles si le modèle est correctement documenté

Modèle climatique, mécanique des fluides.

Données dérivées ou compilées

•Traitement ou combinaison de données« brutes »,

•Reproductibles, mais coûteuses

Base de données compilées, fouille de texte

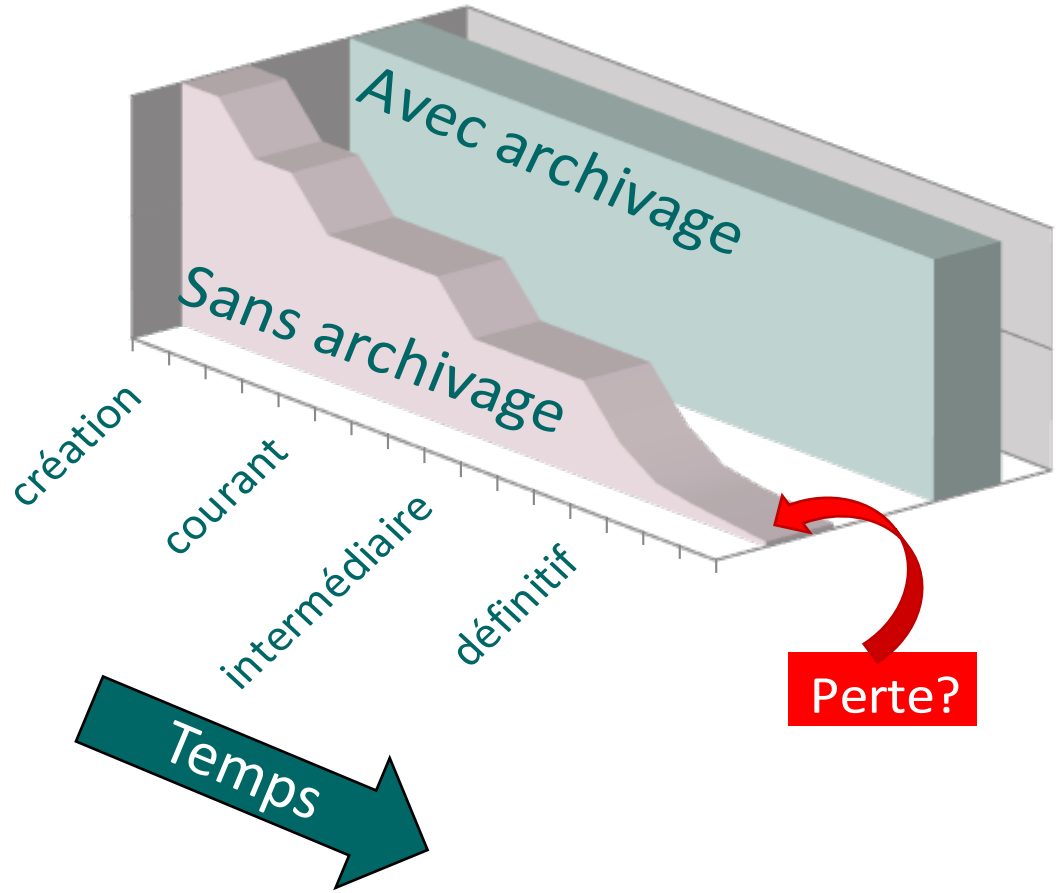
Données de référence

Séquences de gènes, structures chimiques, etc.

EMBL European Molecular Biology Laboratory

International Cancer Genome Consortium

La problématique



L'objectif : conserver le document et l'information qu'il contient :

- Dans le fond et la forme ;
- Sur le très long terme ;
- En le rendant accessible.

Les risques inéluctables :

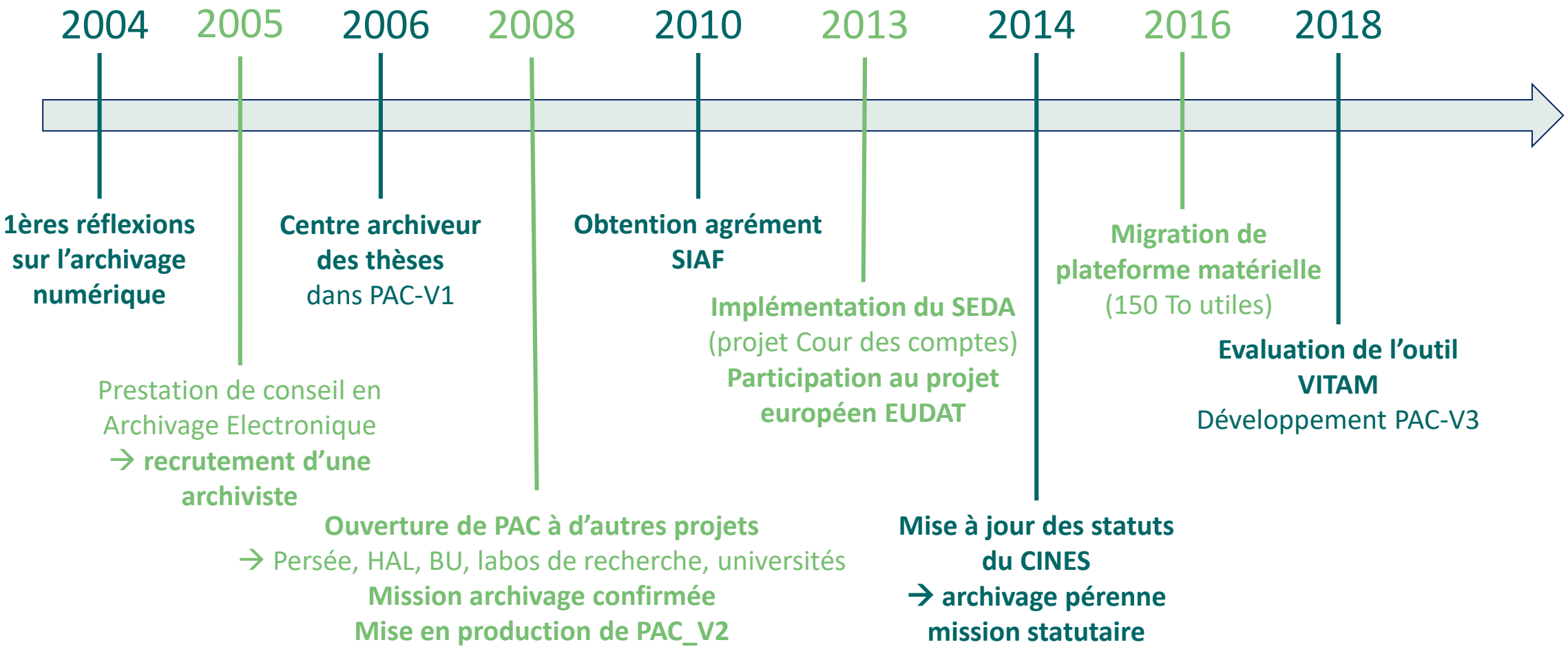
- Connaissance perdue du contenu des fichiers ;
- Format de fichier inconnu ;
- Support physique détérioré ;
- Logiciel ou matériel de lecture disparu.

Les défis

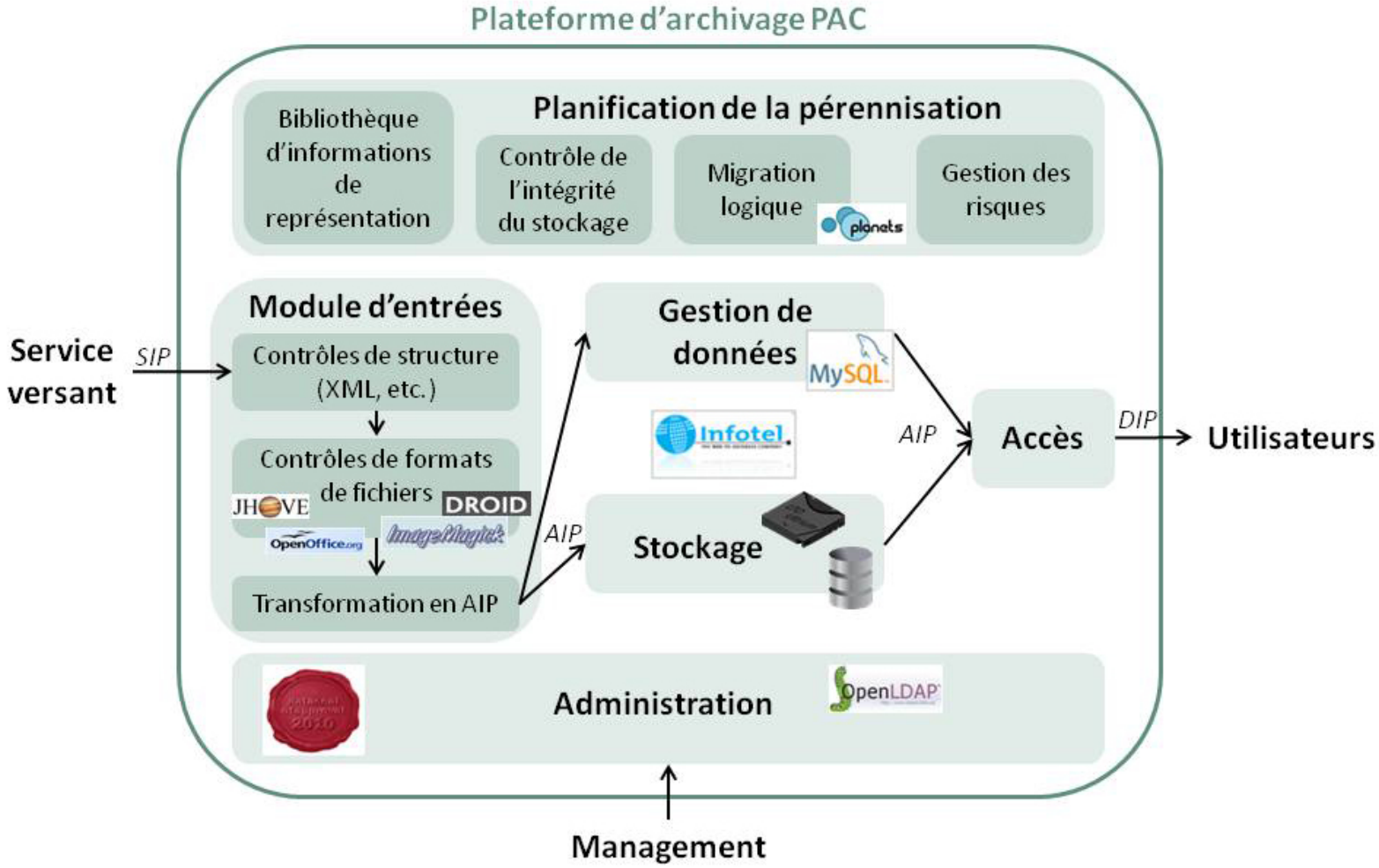
Mise en place de procédures d'assurance qualité pour atténuer l'impact des risques lorsqu'ils se réalisent

Contrainte	Solutions
Connaissance du contenu	• Utilisation de métadonnées • Identification unique et pérenne des documents archivés
Format de fichier inconnu	• Privilégier les formats durables • Identification, validation des formats • Migration logique (conversion de formats)
Support physique détérioré	• Gestion du vieillissement des médias • Migration physique (changement de support)
Logiciel ou matériel de lecture disparu	• Veille technologique et anticipation

L'historique



L’architecture



Le processus d'archivage

1. Réception
- Authentification LDAP
- Transfert SIP
- Accusé de réception

2. Contrôle qualité
- Contrôle checksum
- Contrôle forme du SIP

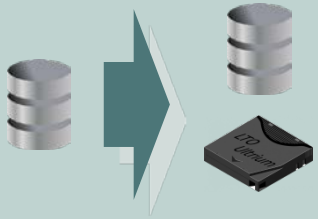
3. Création de l'AIP
- Génération checksum
- Date d'archivage
- PID (ARK, DOI, Handle…)

4. Traitements complémentaires
- Information de représentation

5. Stockage de l'AIP
- Copies multiples

6. Vérification périodique des AIPs
- Contrôle checksum
- Détection corruption
- Validation référenciel

7. Migration physique



8. Migration logique

Plateforme d'archivage pérenne*(SIAF + CoreTrust Seal + ISO 16363)*

La qualité des métadonnées

Préservation des informations décrivant les objets numériques :

- Métadonnées / informations de pérennisation (descriptives, source, historique) ;
- Métadonnées / informations de représentation (techniques, structure).

Plusieurs contrôles de la qualité peuvent être effectués :

- Contrôle du format de la métadonnée par l'adoption d'un standard
  - Métadonnées génériques pour la description des ressources numériques : ex. Dublin Core (ISO 15836) ;
  - Métadonnées spécifiques à un domaine : ex. commerce électronique ebXML, données géographiques (ISO 19115) ;
  - Métadonnées techniques : préservation (PREMIS, METS), propriété intellectuelle (indecs, MPEG-21).
- Contrôle de la valeur des métadonnées par une logique applicative métier
  - Liste de valeurs autorisées, etc.

La qualité des formats de fichier

- **Format maîtrisé = identifié et vérifiable**
  - Format publié ; ex. WAVE, SVG ;
  - Format largement utilisé ; ex. XML, MPEG4 ;
  - Format normalisé si possible ; ex. PDF (ISO 32000-1:2008), PNG (ISO 15948:2004).
- **Respect des spécifications du format**

Les outils libres Jhove, ImageMagick, DROID, ODF Validator permettent une identification, validation et caractérisation des formats.

Type	Format
Text	HTML, PDF, TXT, XML, ODT
Picture	GIF, JPEG, TIFF, PNG, SVG, JPEG200
Audio	WAV, AIFF, AAC, OGG (VORBIS)
Video	MPEG4, OGG (THEORA), MKV

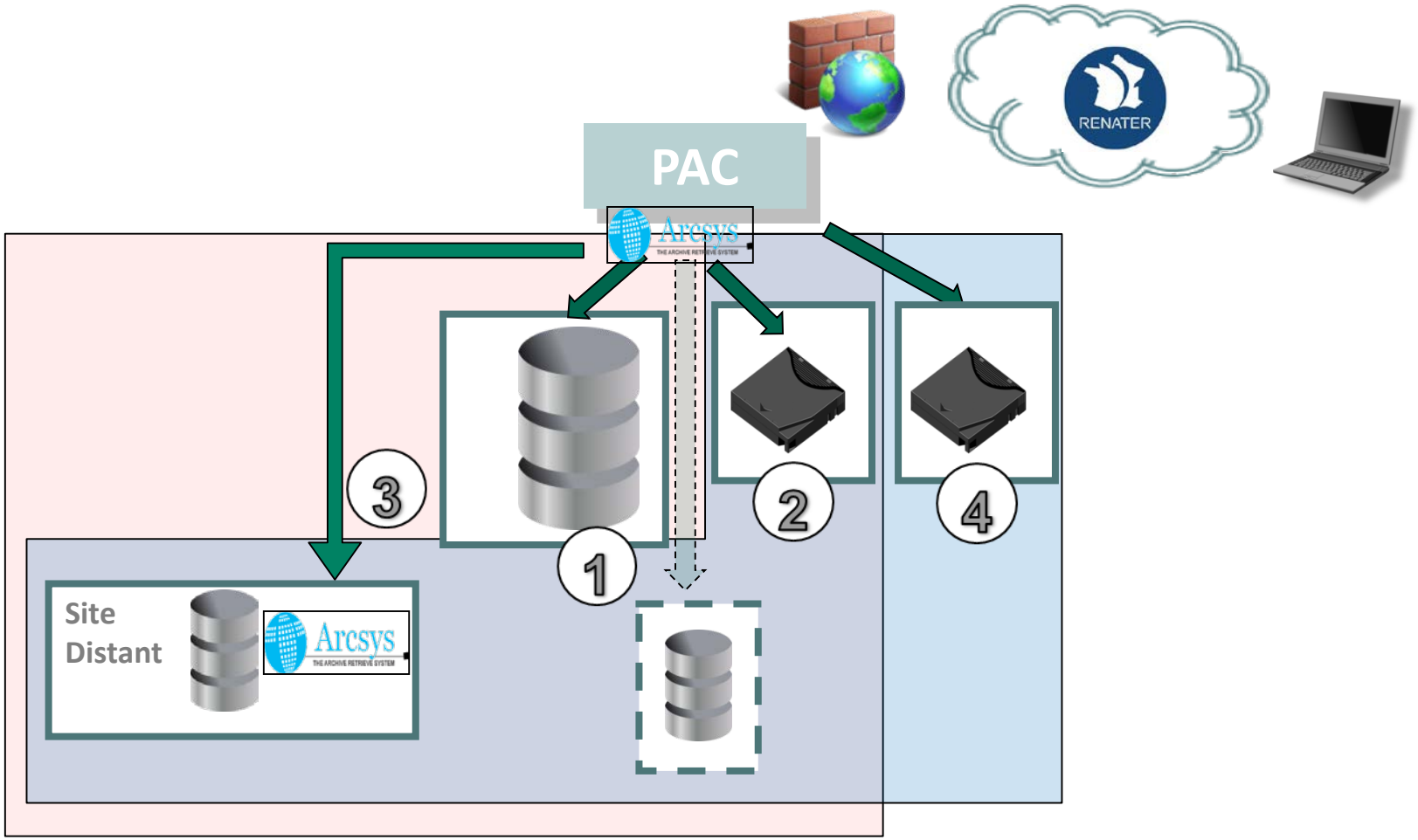
La qualité du stockage

Niveau 1

- 2 copies locales (disque+ bande)
- 1 copie sur site distant

Niveau 2

- 2 copies locales (bande)
- 1 copie sur site distant



Les partenaires

Le projet d’archives

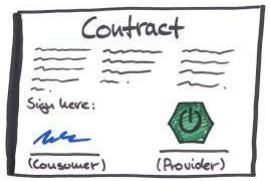
- **Un partenariat encadré :**
  - Lettre d’intention
  - Convention d’archivage
  - Tarification au Téraoctet utile archivé et en fonction du niveau de service

- **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

Le modèle économique

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE
DE L'ÉDUCATION NATIONALE,
DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE



CINES

Subvention d'opération

Contribution financière

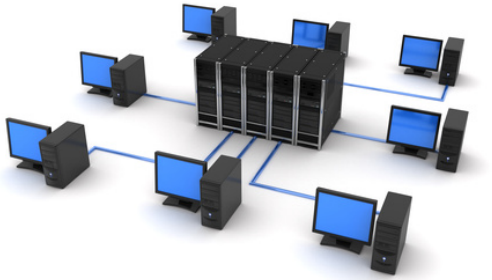
La sécurité des données

- **Protection contre les coupures de fluides**
  - Miroir électrique
  - Réserves d'eau
- **Protection du site : mise en sûreté**
  - Accès protégées par badges / clefs
  - Pièce d'identité + accès physiques tracés + visites guidées
  - Vidéo surveillance du périmètre + système d'alarmes
- **Protection des accès aux systèmes informatiques**
  - Mots de passe robustes,
  - Filtrage des adresses IP ,
  - Validation par HFDS de toute ouverture de compte
- **Protection des fichiers et des données**
  - Protection des accès, cryptage, sauvegarde,
  - Archivage pérenne (DSA, agrément SIAF)

L’infrastructure

- **Infrastructures sécurisées**
  - 5 salles machines : 1 400 m²
  - Locaux techniques : 2000 m²
  - 2 lignes ERDF pour un total de 12,5 MW
  - Données en double alimentation ondulée et sécurisée par groupe électrogène
  - Copies et sauvegardes dans des salles distinctes + copie à distance

- **Ressources**
  - ✓Deux supercalculateurs de niveau mondial
  - ✓Capacités de stockage de plusieurs PetaOctets
  - ✓Des accès réseau performants
  - ✓Des équipes d’experts



Les perspectives

Au plan national, le CINES est maintenant un des acteurs principaux du domaine de l'archivage pérenne

- Rôle étendu dans la stratégie nationale de préservation du patrimoine numérique scientifique de l'ESR
- Impliqué dans de nombreux groupes de travail ou initiatives nationales ou européennes
  - France : PIN ; Europe : CoreTrust seal, RDA-France, etc.

<u>Objectifs 2019-2020 :</u>

- Déploiement de PAC-V3 (Q1/2019)
- Préservation de données de santé (2019)
  - Partenariat avec l'INSERM
  - Projet France Médecine Génomique 2025
- Amélioration continue de l'assurance qualité (2019/2020)
  - Renouvellement des certifications CoreTrust seal et agrément SIAF (2019)
  - Audit pour certification ISO (2020)
- Migration vers VITAM (2020)

CORE TRUST SEAL

PROGRAMME vitam archivage numérique

Des questions ?

CONTACTS

Olivier ROUCHON
Resp. DAD – CINES
rouchon@cines.fr