ARCHIVAGE INTERMÉDIAIRE DE DONNÉES SCIENTIFIQUES

ISAAC

INFORMATION SCIENTIFIQUE ARCHIVÉE AU CINES

Le 04 janvier 2013

PIN - Pérennisation et communication de l'Information Numérique

Basé à Montpellier (Hérault, France)

EPA créé en 1999, succédant au CNUSC (Centre National Universitaire Sud de Calcul) – créé en 1980

Sous tutelle du Ministère de l'Enseignement Supérieur et de la Recherche

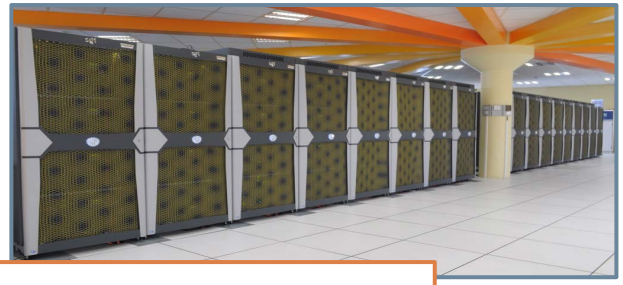
- DGRI (Direction Générale de la Recherche et de l'Innovation)
- DGESIP (Direction Générale pour l'Enseignement Supérieur et l'Insertion Professionnelle)

Missions

- Calcul numérique intensif (30 ème machine mondiale)
- Archivage pérenne de documents électroniques
- Activité transversale : hébergement d'environnements informatiques

plus d'info sur : http://www.cines.fr

Les données calcul *(projet d'un an renouvelable)*



Étape 1 : Dépôt
Codes, Données calcul

700 To d'espace de travail optimisé pour le calcul

Lancement de job,
Ecriture des fichiers de sorties,
Compilation, dépôt de librairies…

Étape 2 : transfert des données pertinentes (Valeur ajoutée 237 TéraFlops)

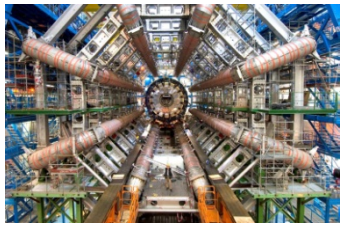
600 To de stockage sauvegardé sur la durée du projet+6mois 600To

HSM

Robothèque

Panorama de la donnée Scientifique

- 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 recherches 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

BIG DATA

BIG DATA n'est pas à priori un phénomène de mode mais une évolution de l'organisation des Systèmes d'information ***qui nécessite une architecture en dehors des outils conventionnels de gestion.***

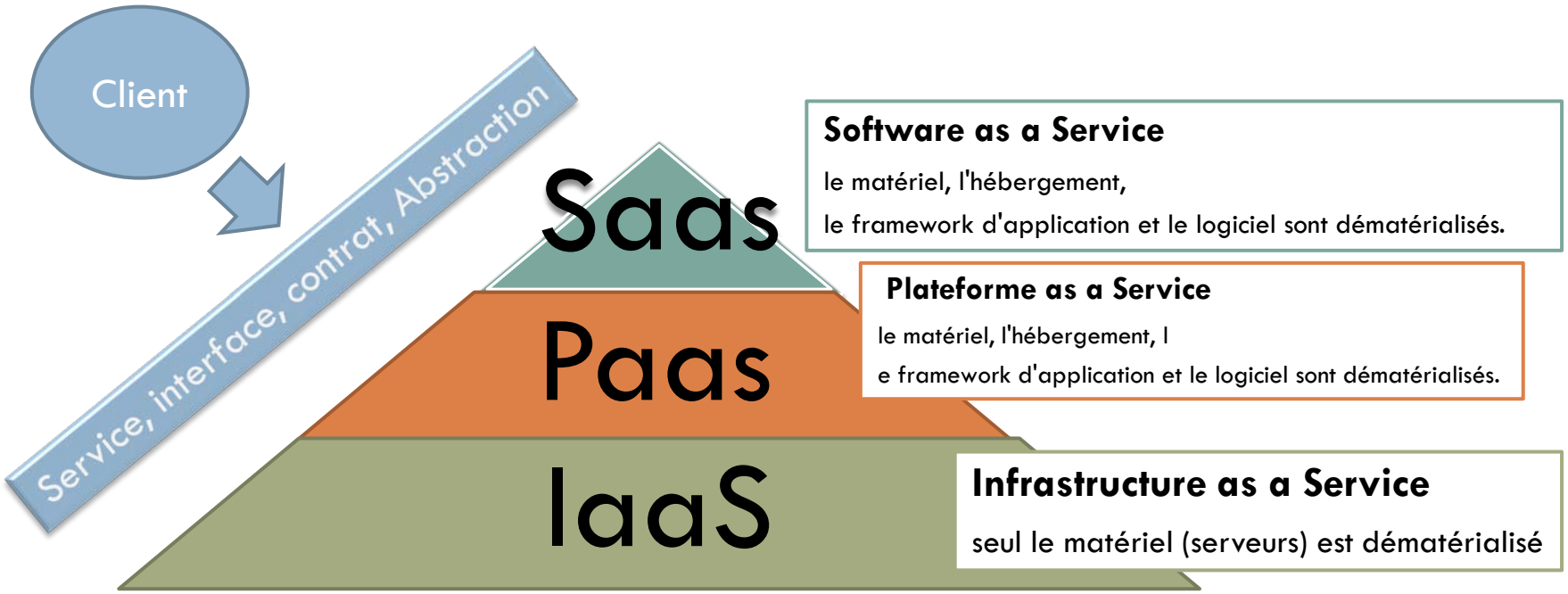
Comment

- Stocker : niveau de sécurité, délais d'accès, autorisation, authentification
- Pérenniser: garantir l'exploitabilité d'une donnée dans le futur.
- Traiter , Analyser : traiter de grande quantités de données hétérogène
- Exploiter, Visualiser : web 2.0 pour exploiter la donnée là ou elle est
- Donner un sens : savoir distinguer la donnée de l'information
- Protéger : donnée (partiellement) publique, privée, répartie, niveau de confiance
- supprimer : supprimer une donnée qui à de multiples copies, supprimer le train de bit.

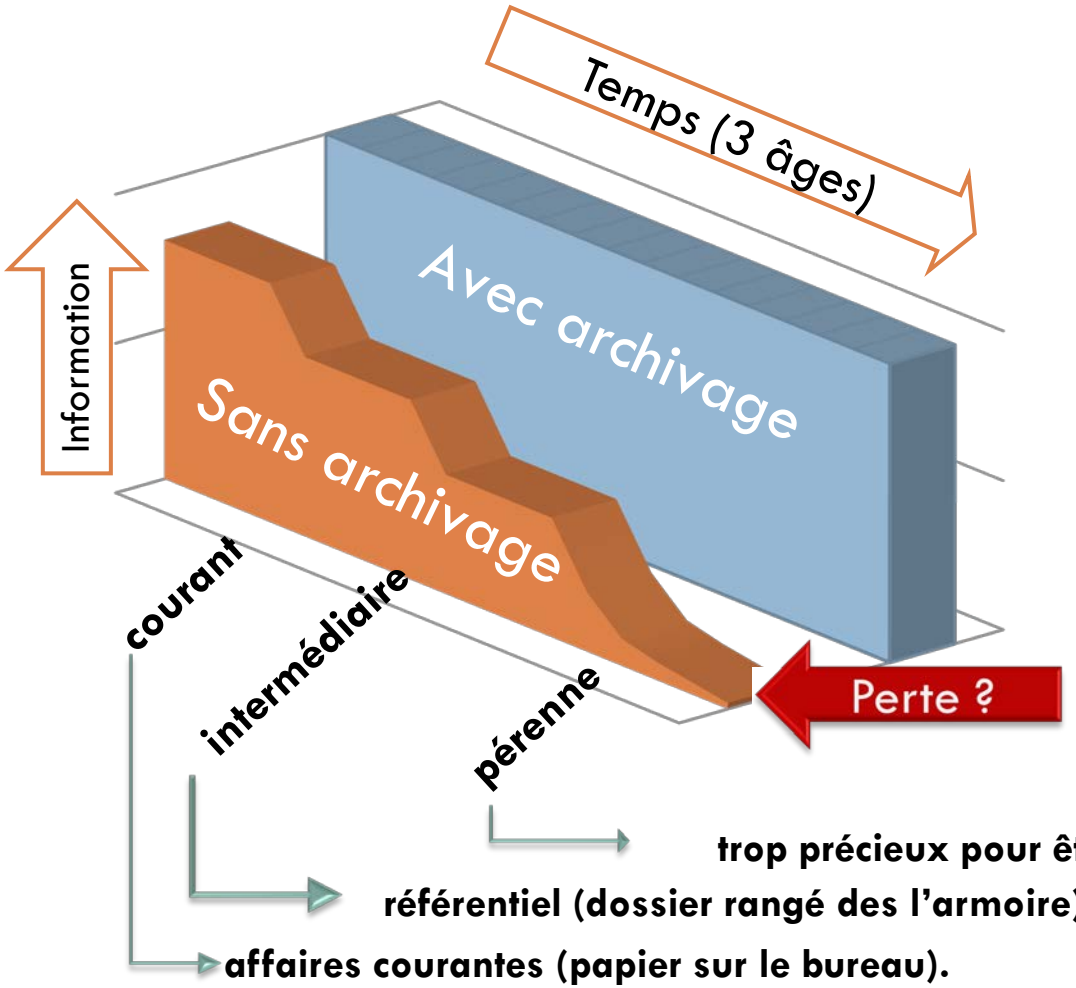
Des réponses à plusieurs niveaux

- Politique : financement, privé, publique
- Organisationnel : communautés scientifiques, laboratoire, chercheur isolé
- Méthodologique : protocoles d'accès, d'intérogation, représenter l'information
- Algorithmique : mapReduce, indexation,
- Interopérabilité : standard de formats, de description, d'accès
- Applicatifs : outils sur poste client, sur serveur, Web 2.0
- Infrastructure : gestion du risque cout/sécurité – délais de réponse

CLOUD


Client
Service, interface, contrat, Abstraction
Saas
Paas
IaaS
Software as a Service
le matériel, l'hébergement,
le framework d'application et le logiciel sont dématérialisés.
Plateforme as a Service
le matériel, l'hébergement, l
e framework d'application et le logiciel sont dématérialisés.
Infrastructure as a Service
seul le matériel (serveurs) est dématérialisé

Les risques lié à l'information



Risque sur :

- **Compréhension**
- **Intégrité**
- **Exploitation**
- **Valorisation**

Mise en œuvre

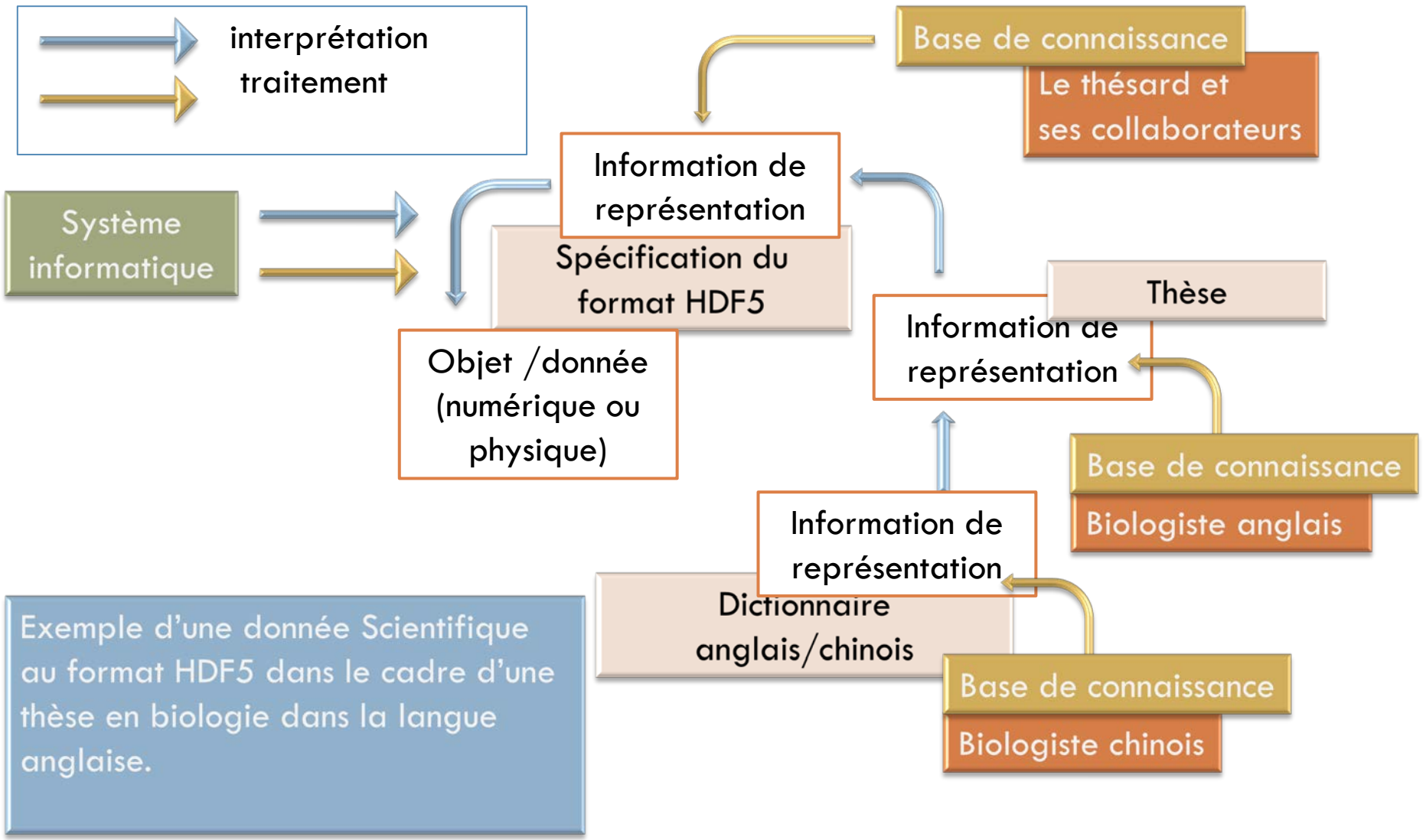
- Métadonnées
- Contrôle des formats
- Communauté structurée
- Contrôle d'intégrité
- Veilles

trop précieux pour être détruit (au grenier).

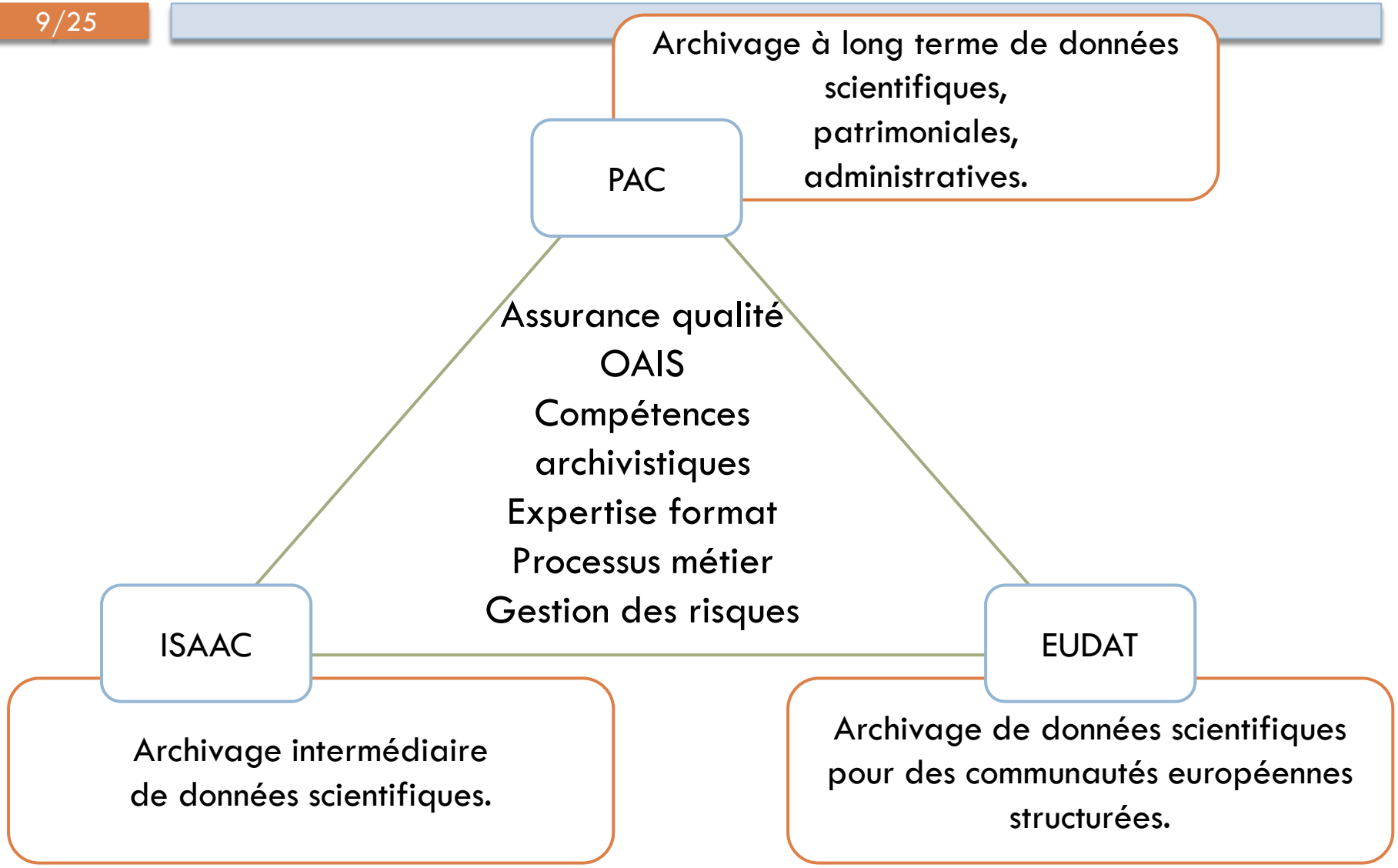
référentiel (dossier rangé des l'armoire).

affaires courantes (papier sur le bureau).

Donnée et Information



L'archivage au CINES



Les systèmes d'archivages du CINES

PAC

- **archivage intermédiaire et à long terme de données administratives, patrimoniales et scientifiques**
  - Mandat pour l'archivage des thèses électroniques soutenues en France (arrêté du 7 août 2006)
  - Agréé pour l'archivage intermédiaire par le SIAF
  - Périmètre opérationnel : données de l'enseignement supérieur et de la recherche
  - Partenariat avec le TGE Adonis : archivage et diffusion des données numériques en SHS

ISAAC (en développement)

- **archivage intermédiaire de données scientifiques**
  - Dimensionné pour des petites structures ayant de grands volumes de données
  - Une donnée organisée et validée par des communautés d'experts
  - Un travail scientifique valorisé par le partage et la diffusion

Organigramme « *Fonctionnel* »

Département Archivage et Diffusion

6 Ans d'experience

- Compétences Archivistique, Normes et standard
- Processus métier, Droit informatique
- Connaissance des institutions, veille technologique
- Expertise en format de fichier et métadonnées

Des Outils

- **applicatifs :** contrôleur de format
- **conceptuels :** modèle OAIS…..
- **méthodologiques :** processus métiers

Département Calcul Intensif

- Étude, optimisation des codes de calcul scientifique
- gestion de grand volume de stockage
- formats de données scientifiques
- Support utilisateur
- Expertise sur les dossiers de demande de ressources

Département Services Informatiques et Infrastructures

serveurs, HSM, robothèque, réplication distante, réseau Renater

Opportunité

Expertise

- Archivage pérenne
- calcul, visualisation scientifique
- Format
- Stockage

Atouts

- Relation avec un grand nombre de laboratoires
- Implication dans les projets Nationaux et européens (Eudat / Prace)
- Capacité d'organiser et de fédérer des communautés scientifiques

Des challenges

- Sensibiliser et impliquer les utilisateurs des communautés scientifiques
- Proposer des services proches des besoins des utilisateurs pour susciter l'adhésion
- Mettre en œuvre des moyens mutualisés répondant
- aux normes et standards
- aux besoins des utilisateurs

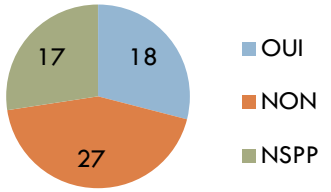
Objectif

Mettre en place un service d'archive intermédiaire de données scientifiques pour une communauté d'utilisateurs structurée partageant les même formats de données et désireuse de mettre en œuvre une démarche combinant diffusion et conservation à long terme.

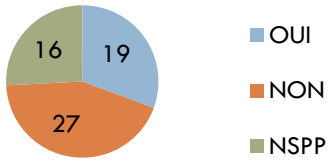
- Démarré en juin 2010.
- RH : globalement 2ETP
  - responsable projet mi temps
  - développeur plein temps
  - expert calcul scientifique
  - infrastructure, administration, web

ENQUETE auprès de laboratoires calculant au CINES (155)

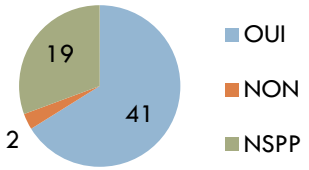
Sensible à l'archivage



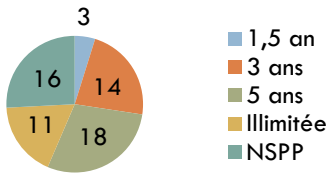
Solutions mises en œuvre ?



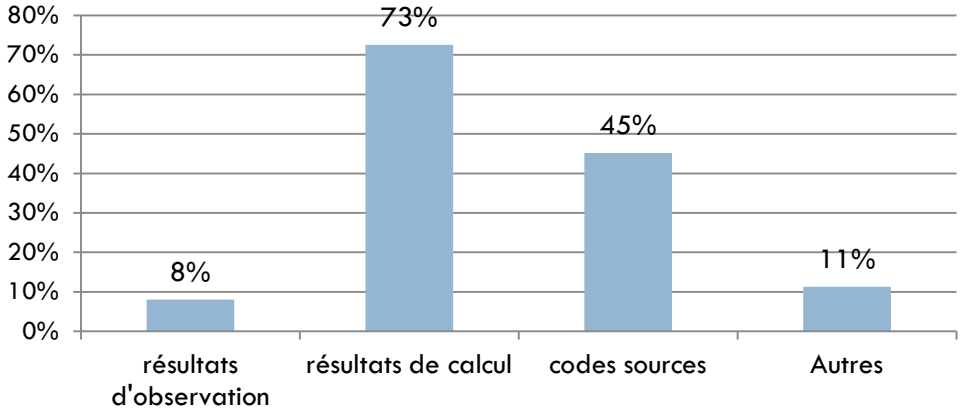
Besoin d'informations ?



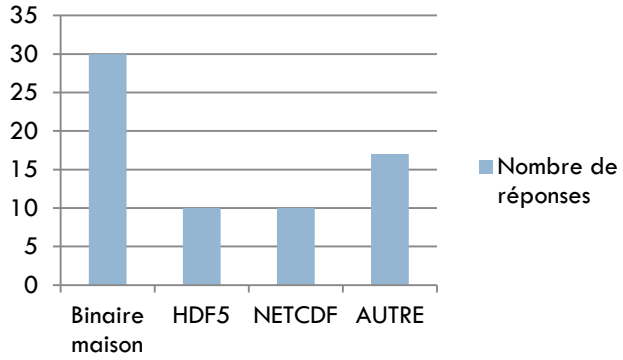
Durée de stockage souhaitée



Besoins selon le types de données à archiver



Formats utilisés



Constat

- Le stockage dans les laboratoires se fait sur des systèmes locaux
- Importance de la sécurisation des données
- Besoins d'accès le temps de l'analyse et de la publication des résultats (de 3 à 5 ans)
- Intérêt de l'archivage pour
  - la reprise de travaux
  - comparaison de résultats
  - post processing
  - partage de données entre scientifiques
- beaucoup de formats binaires « maison » accompagné d'une description « maison »
- La notion même de métadonnée est rarement connue
- Partage de la donnée en cours de travaux dans un cercle restreint de collaborateurs connus
- Volonté de partage à un communauté plus large (après publication).
- Volumes de données par projet de 1 à 10 To globalement.

Bilan

- Besoin d'information sur les enjeux de l'archivage et de la diffusion
- Une solution par laboratoire -> besoin de fédération
- Conserver l'information n'est pas le métier des laboratoires
- Exploiter et publier les résultats : 3 à 5 ans
  - Principalement des résultats de calcul
  - post-processing et comparer des résultats,
  - ne pas refaire des calculs couteux
- Partager et sécuriser la donnée dans un cercle restreint

Cerner l'importance d'une donnée est une démarche à part entière.
Il est nécessaire de mettre en œuvre les processus permettant
Son partage, sa valorisation, sa conservation et son intégrité.

Une organisation

Le CINES identifie des Comités Thématiques Archivage (idem DARI)

- Président
- Groupe d'expert de la discipline
- Référent CINES

Le CTA propose des critères pour les projets qu'il contiendra

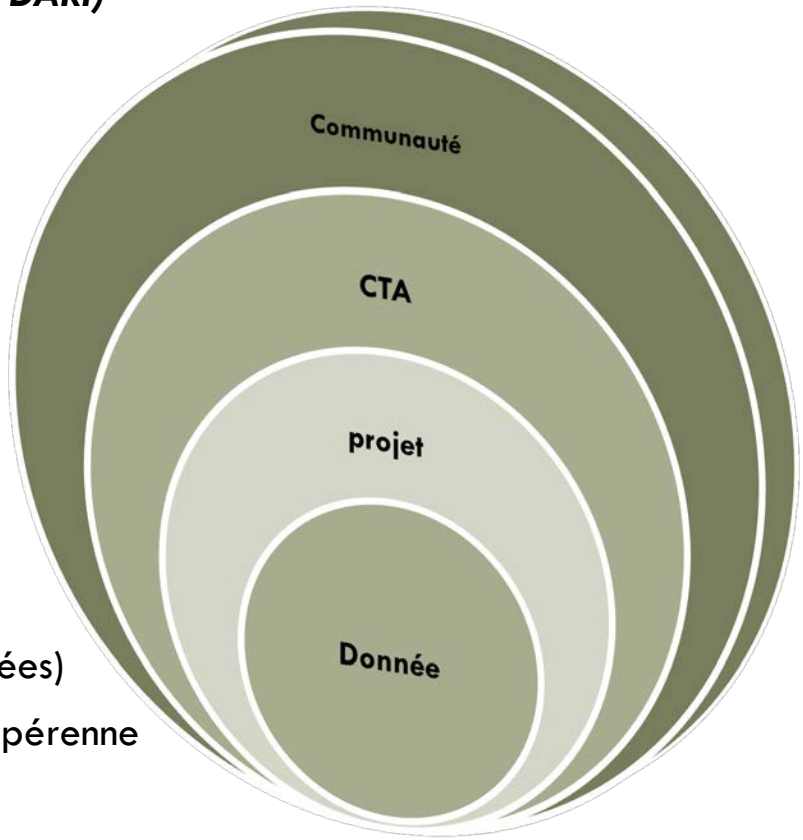
- Jeux de métadonnées
- Liste de formats
- Autorise les projets en accord avec le CINES

Le projet

- Transmet les données à conserver
- Remplit les conditions du CTA (format et métadonnées)
- Récupère les données, ou migre vers un archivage pérenne

Communautés

- Accèdent aux informations selon leurs autorisations



Les étapes de la vie d'un projet d'archivage dans ISAAC

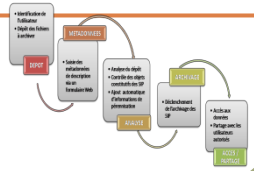
J'ai des données scientifiques à archiver

Je remplis le formulaire Web de demande d'ouverture d'un projet d'archivage de données scientifiques dans ISAAC

Le Comité Thématique d'Archivage de mon domaine scientifique étudie et valide mon projet

Je travaille avec le CINES à la mise en œuvre des éléments spécifiques de mon projet.

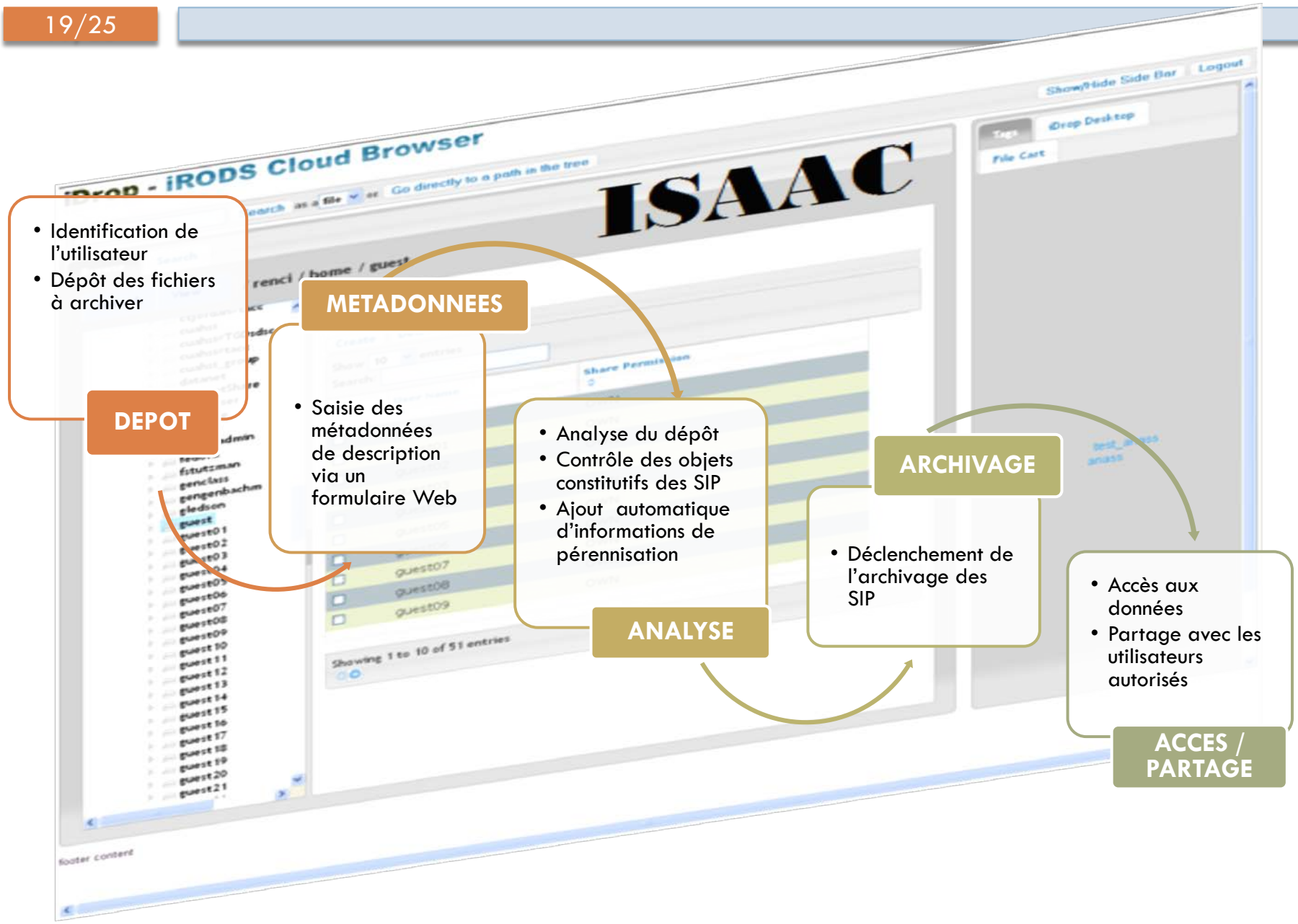
Mon projet entre en production, je réalise l'archivage de mes données sur l'interface ISAAC



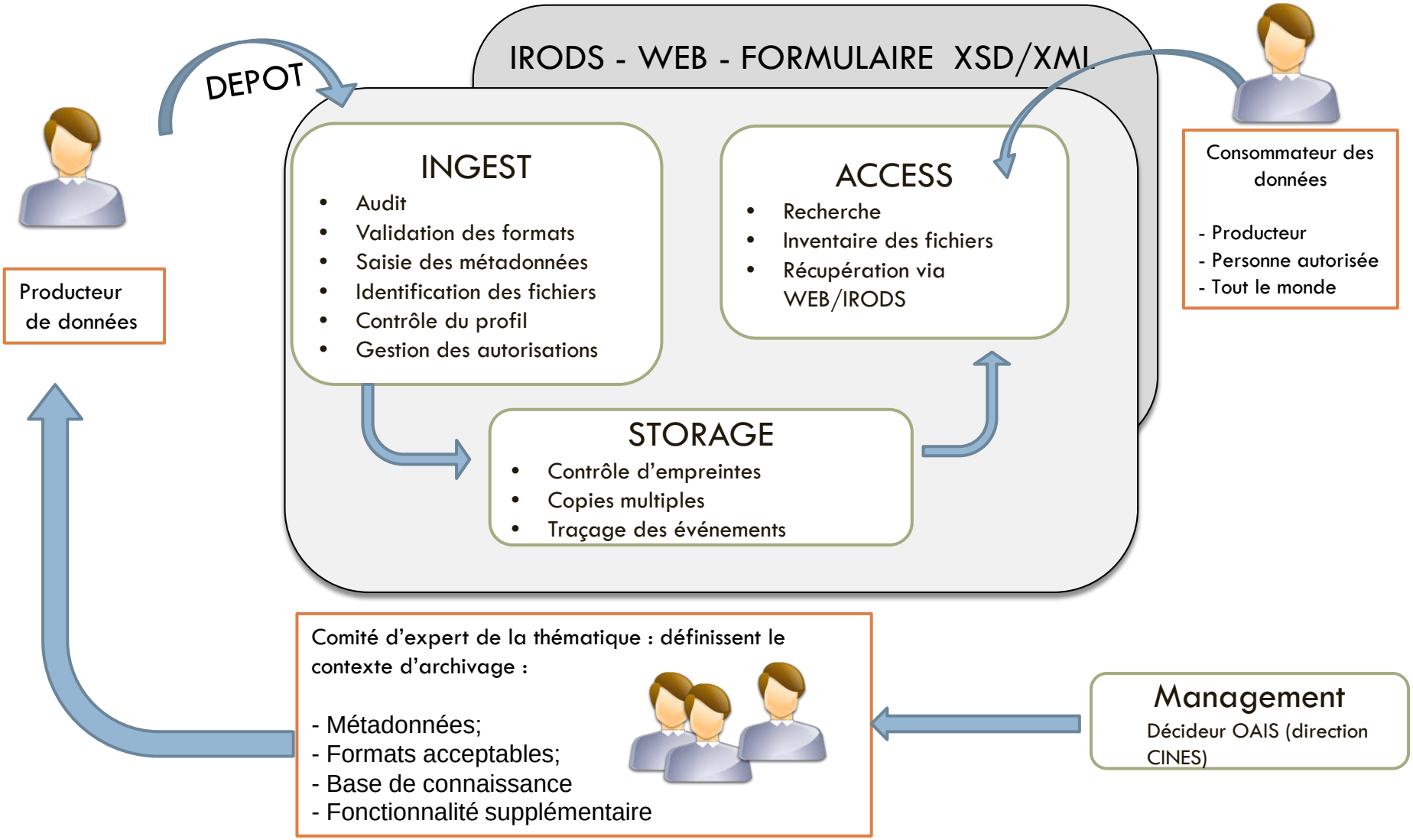
Moi-même, les membres de mon équipe ou d'autres utilisateurs autorisés pouvons consulter et partager les données du projet sur le site d'ISAAC

Au terme de la durée de conservation
Je récupère mes données OU
je lance un processus d'archivage sur le long terme dans PAC (si données éligibles)

Les fonctionnalités de l'interface WEB



Représentation fonctionnelle



Les Technologies utilisées

basée sur des technologies libres de droits : Irods, XML, XSD, SGBD Postgres, AJAX.

Irods

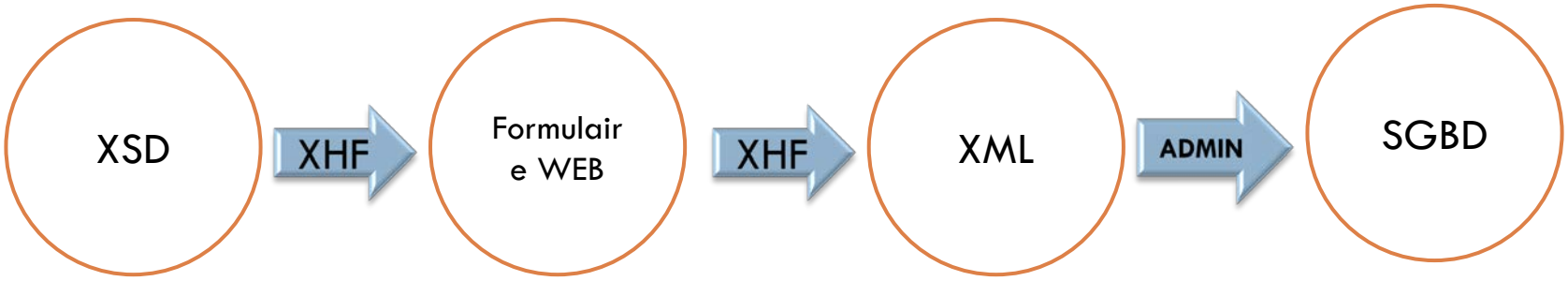
- Données réparties
- Gestion souple de grands volumes
- Règles permettant des traitements spécifiques (calcul de checksum, migration)

Interfaces :

- web 2.0 (ajax), client irods

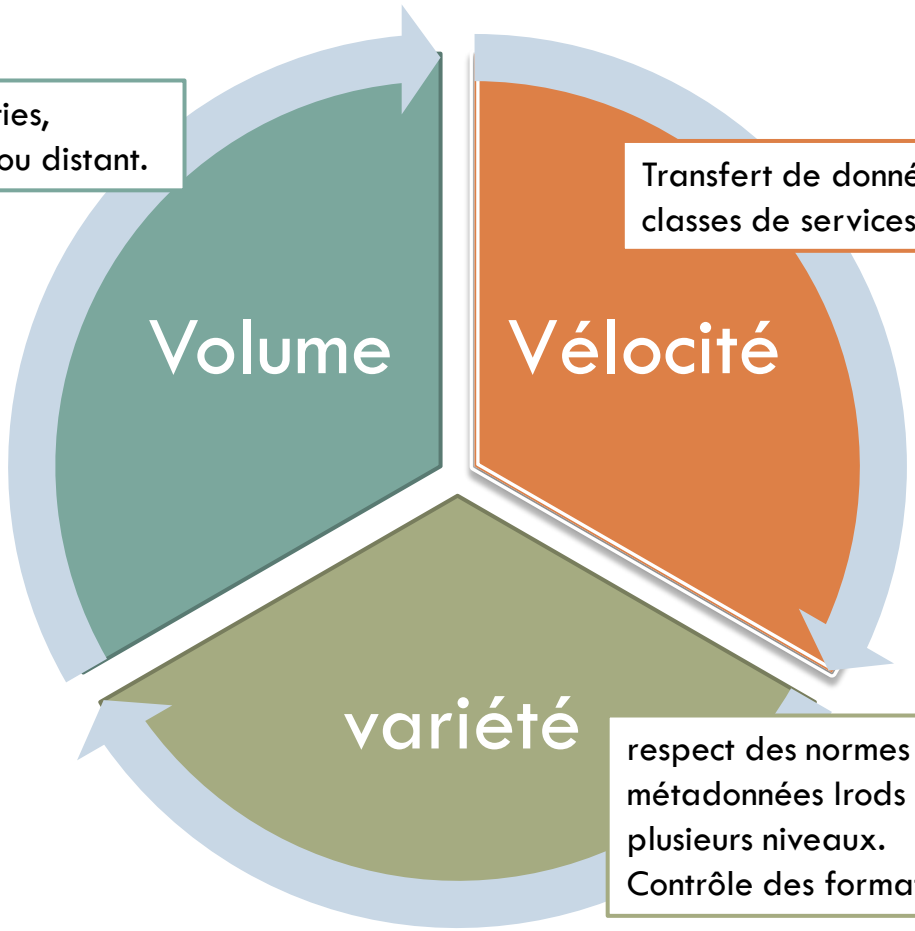
Implémentations

- Validateur de format (hdf5)
- Validateur XML/XSD.
- XHF : génération dynamique de formulaire web à partir de schéma XML



Isaac, prise en compte des préoccupations du Big Data

Données iRODS locales ou réparties, disques, de bandes sur site local ou distant.



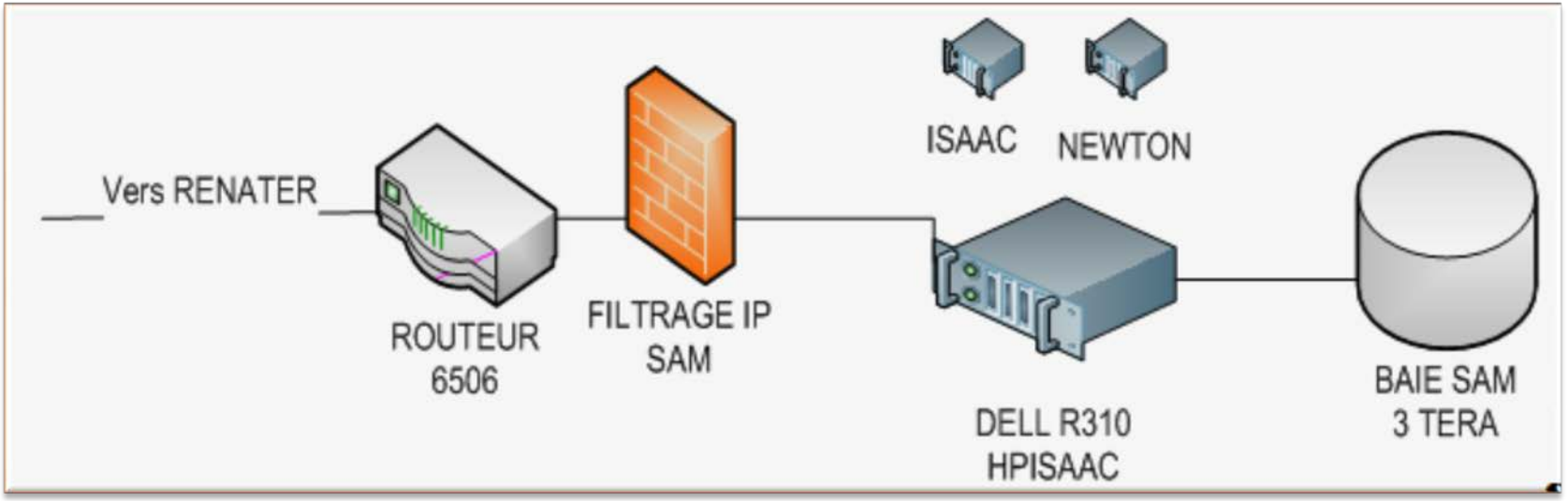
Transfert de données via iRods, classes de services.

respect des normes : OAIS, NFZ-42-013, métadonnées Irods et métadonnée (XML) à plusieurs niveaux.
Contrôle des formats

Infrastructure matériel

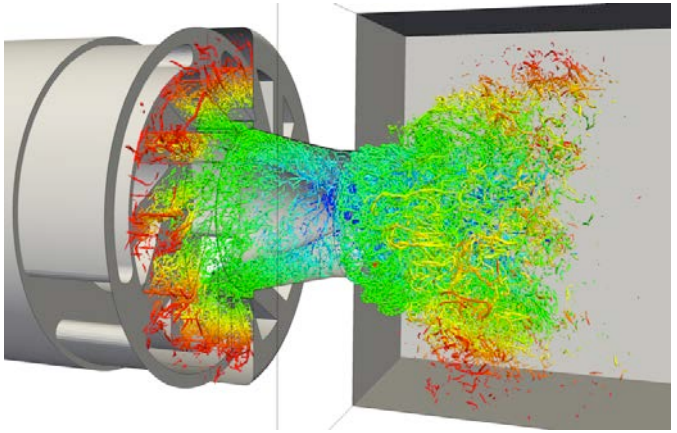
Environnement PROD : 2 serveurs virtuels hypervisé par XEN sur un DELL POWEREDGE R310, 4 Go RAM, 133 go Disque.

- Newton : 6 UC virtuelles, 1Go mémoire,47 Go, 100 Go sur SAM. (test pre-archivage client (PRECCINSTA+XHF) newton.cines.fr)
- Isaac : 6 UC virtuelles, 2Go mémoire,47 Go, 3To sur sam. (ISAAC, pilote)
- Sauvegarde TSM prévue dans l'attente d'une configuration iRods.



Projet Pilote PRECCINSTA

- Partenariat avec le CORIA COmplexe de Recherche Interprofessionnel en Aérothermochimie (l'UMR 6614)
- PRECCINSTA : Prediction and control of combustion instabilities for industrial gas turbines
- Volumetrie de 1 à 10 To.



- Exemple :
  - Un jeu de de données contient :
  - 4096 fichiers maillage au format HDF5
  - 4096 fichiers solution au format HDF5
  - 4096 fichiers de méta-données associés à chaque morceau du maillage + 1 fichier de méta-données "maitre", le tout au format XDMF afin de pouvoir importer les solutions dans paraview.

Merci pour votre attention

?

prat@cines.fr