Archivage intermédiaire de données Scientifiques

ISAAC –

Information Scientifique Archivée Au Cines

Constat

Enquête auprès des laboratoires de recherche

- 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 et il est nécessaire de mettre en œuvre les processus permettant sa valorisation, sa conservation et son intégrité

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 à relever

- 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.

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

Infrastructure-technologies

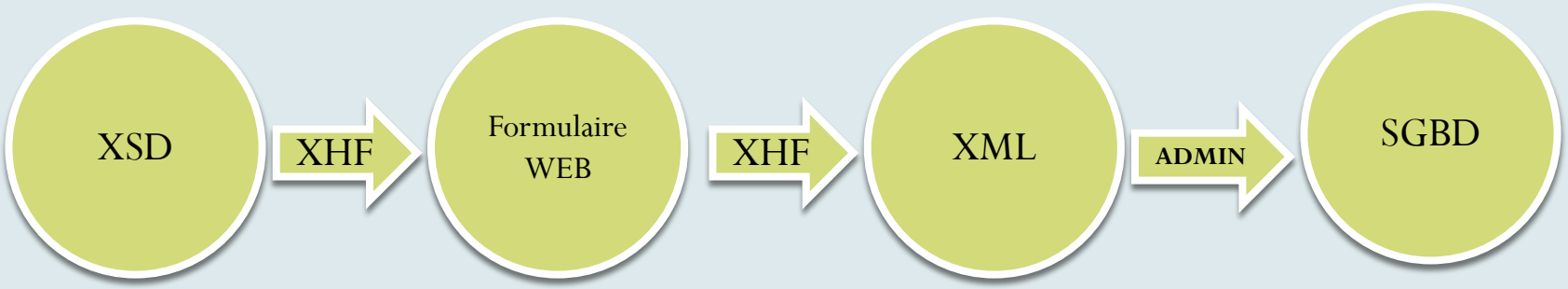
Irods

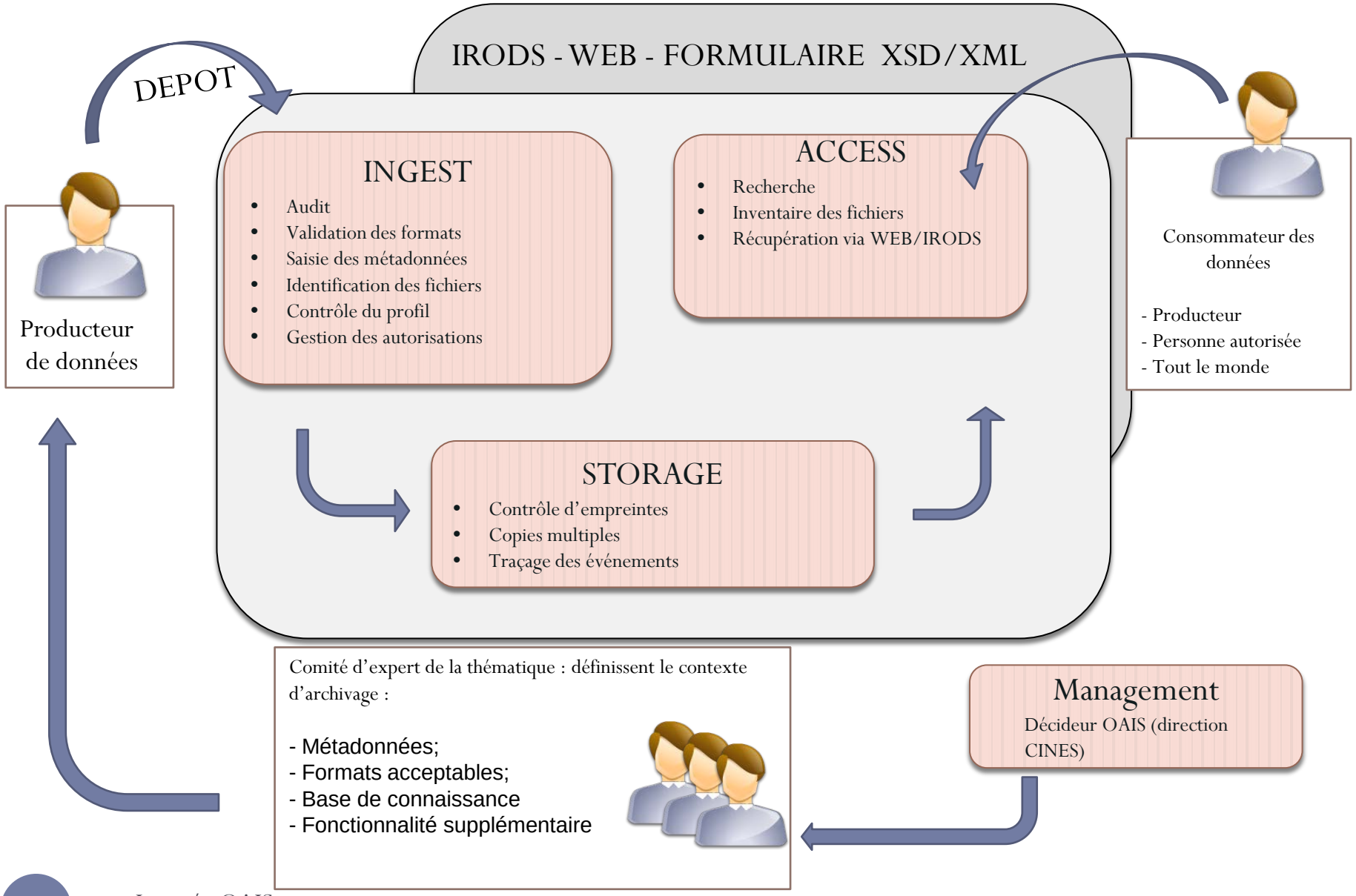
- Données réparties
- Gestion souple de grands volumes
- Règles permettant des traitements spécifiques (calcul de checksum, migration)

Interfaces : web 2.0 (ajax), client irods

Implémentations

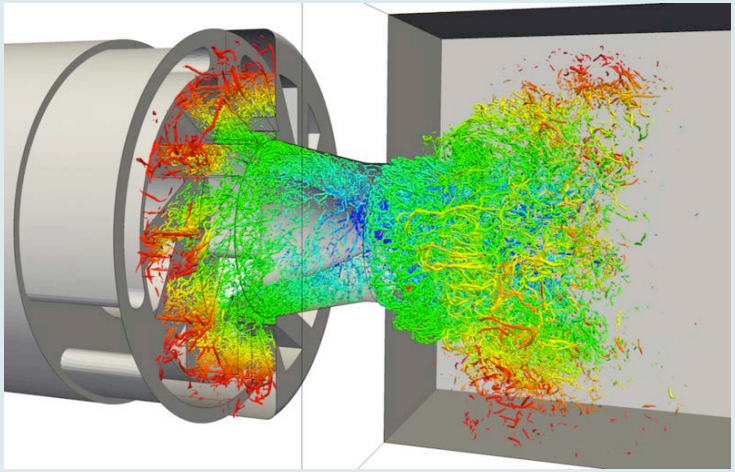
- XHF : génération dynamique de formulaire web à partir de schéma XML




IRODS - WEB - FORMULAIRE XSD/XML
DEPOT
INGEST
Audit
Validation des formats
Saisie des métadonnées
Identification des fichiers
Contrôle du profil
Gestion des autorisations
ACCESS
Recherche
Inventaire des fichiers
Récupération via WEB/IRODS
STORAGE
Contrôle d'empreintes
Copies multiples
Traçage des événements
Producteur de données
Consommateur des données
- Producteur
- Personne autorisée
- Tout le monde
Comité d'expert de la thématique : définissent le contexte d'archivage :
- Métadonnées;
- Formats acceptables;
- Base de connaissance
- Fonctionnalité supplémentaire
Management
Décideur OAIS (direction CINES)

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




?
Contact : prat@cines.fr

Gestion de gros volumes de données

CINES–

Journées OAIS

BIG DATA ?

Quelques chiffres

- ***Production mondiale : 2.5*** exaoctet ***/jours***
- ***90 % des données ont été produites durant les deux dernières années.***

BIG

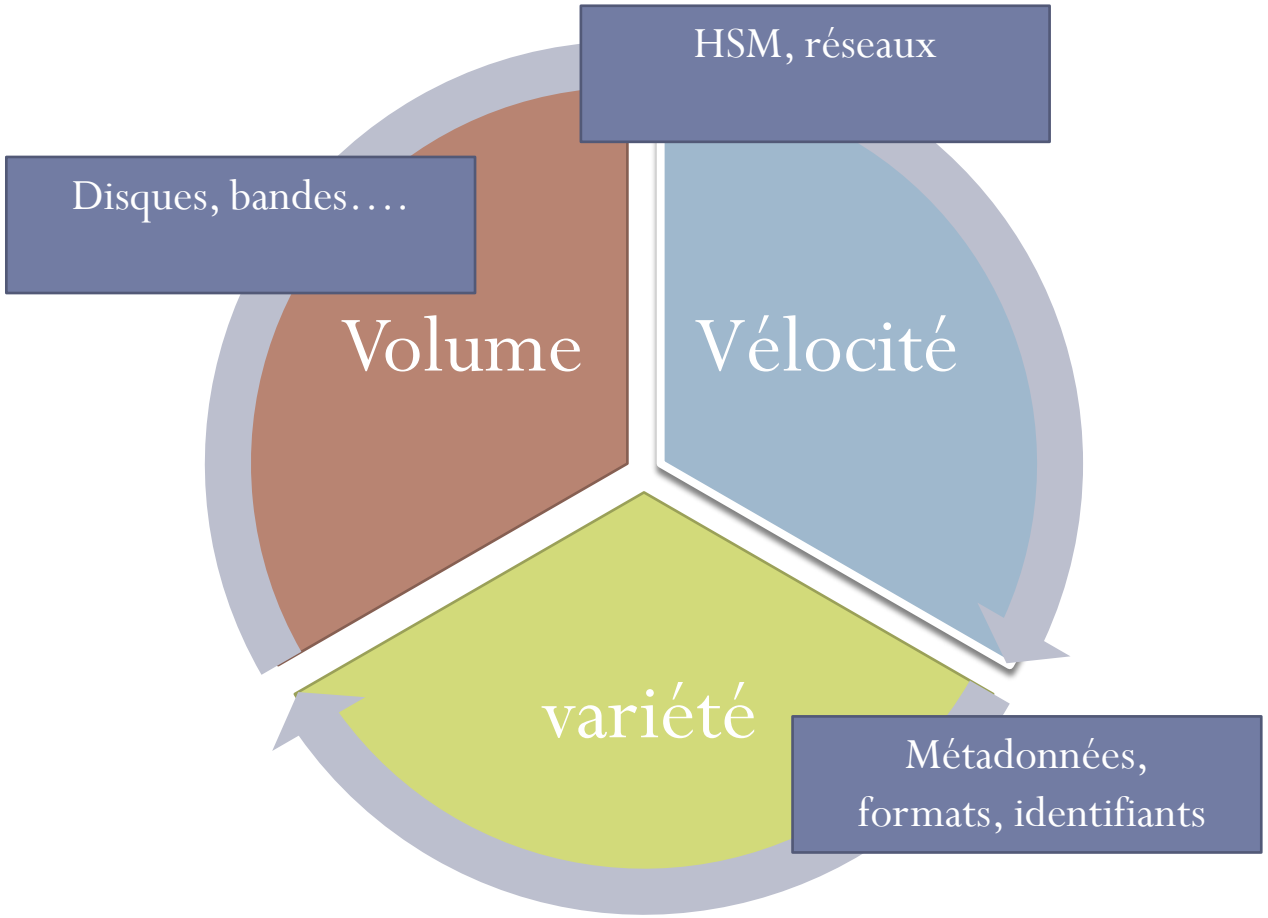
nécessite une architecture en dehors des outils conventionnels de gestion.

DATA

- ***Fichiers***
- ***Base de données***

Le BIG DATA n'est pas à priori un phénomène de mode mais une évolution de l'organisation des Systèmes d'information.

Trois aspects



Stratégie de mise en place d'un SI Big Data

- Mise en œuvre
  - RH : compétences+temps
  - Financements
  - Contraintes liées à la données : sécurité, intégrité, confidentialité
  - Taille de la structure : meso centre, national, européen, international
  - interopérabilité ? Handle, DOI, structuration des données, authentification etc.
- Offre privée : pas de pub ☺
- Initiatives public
  - Eudat : European Data Infrastructure
  - Prace : PartneRship for Advanced Computing in Europe

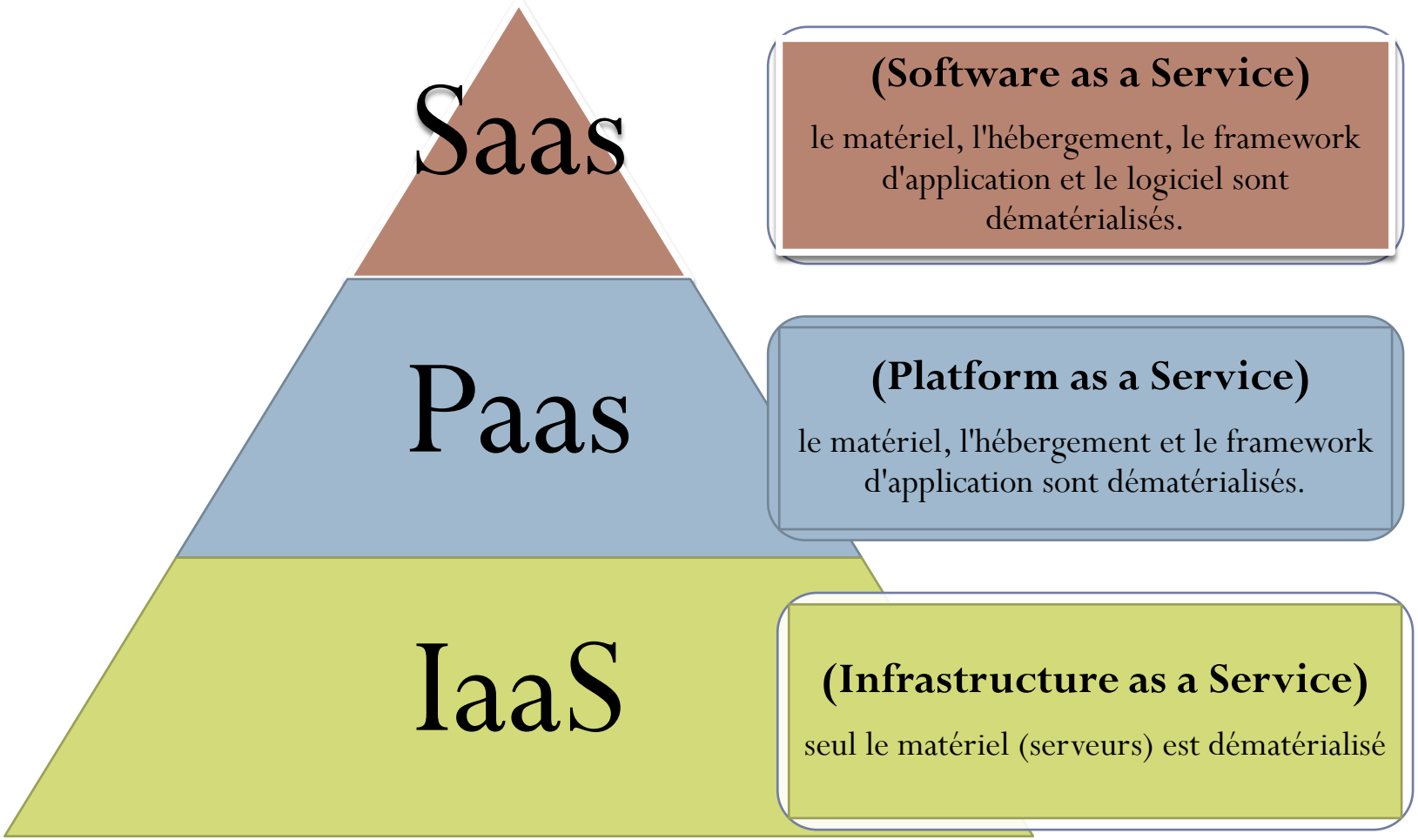
Données privées

Il est impératif de connaître juridiquement à qui appartient l'information.

Directives européennes (95/46/CE) sur l'utilisation des données privées

- **Qualité :** exactes, utile, finalités déterminées
- **Légitimation :** consentement des personnes consternées ou nécessaire pour exécution d'un contrat, contraintes légale, intérêt vital ou public,
- **Hors catégorie** : ethnique, politique, religieux, philosophiques, syndicales, santé, vie sexuelle.
- **Information :** l'identité du responsable des information doit être donnée
- **droit d'accès :** à ses données
- **droit d'opposition :** sur demande et gratuitement, informée si transmise à un tiers
- **confidentialité et sécurité des traitements :** mise en œuvre de mesures appropriées pour protéger les données
- **Notification :** aux autorités de contrôle national

Type de services



ISAAC et le BIG DATA

- **Type de service :** Saas
- **basée sur des technologies libres de droits :** Irods, XML, XSD, SGBD Postgres, XHML, AJAX.

Les 3 aspects :

- **Volume** : gestion par iRODS de grands volumes de données locales ou réparties, ajout facile de disques, de bandes etc..
- **Vélocité** : Protocole Irods pour les échanges de données, classes de services.
- **Variété** :
  - respect des normes : OAIS, NFZ-42-013,
  - métadonnées Irods et métadonnée (XML) à plusieurs niveaux.
  - Contrôle des formats