cnes
CENTRE NATIONAL D'ÉTUDES SPATIALES

Un système générique pour l'accès à des archives de données scientifiques : le SIPAD-NG

Système d'Information, de Préservation et d'Accès aux Données – Nouvelle Génération

Réunion Groupe PIN, 15 janvier 2008

Dominique Heulet (CNES Toulouse)
dominique.heulet@cnes.fr

cnes

Plan de la présentation

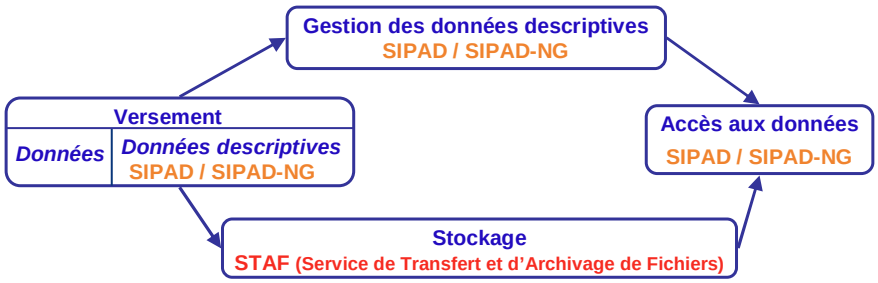
1) Introduction
2) Le CDPP (Centre de Données de la Physique des Plasmas)
3) La généricité dans le SIPAD-NG
4) Démonstration
5) Architecture du système SIPAD-NG
6) Conclusion

Origine des systèmes SIPAD et SIPAD-NG

- **SIPAD : système d'accès à l'archive du CDPP**
  - **Accès Web au catalogue des données du CDPP – Fonctions de :**
    - Recherche / sélection de données par navigation, critères de sélection, imagettes
    - Commande des données sélectionnées et récupération :
      - sur un espace utilisateur,
      - par transfert FTP vers une machine de l'utilisateur,
      - par média (CD, DLT)
  - **SIPAD opérationnel pour le CDPP de 1998 à 2007**
- **SIPAD-NG :**
  - **Remplace depuis mars 2007 le SIPAD pour l'accès aux données du CDPP**
  - **Offre les mêmes fonctions que le SIPAD + des améliorations :**
    - Traitement des données par des SVA (Services à Valeur Ajoutée)
    - Interfaçage avec des applications externes (interopérabilité)
    - Choix techniques complètement différents

Situation par rapport au modèle OAIS

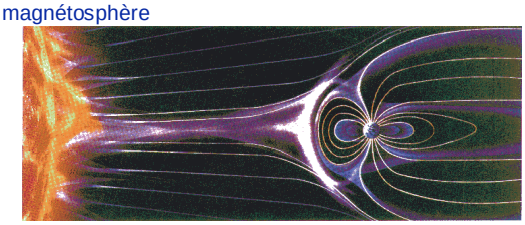
- **En plus des fonctions d'accès aux données, les systèmes SIPAD et SIPAD-NG offrent des fonctions d'acquisition et de gestion des données descriptives :**
  - **Métadonnées, imagettes, documents associés aux données**



- **Les systèmes SIPAD et SIPAD-NG ont été conçus pour être utilisés par d'autres thématiques que la Physique des Plasmas**

Présentation du CDPP (Centre de Données de la Physique des Plasmas)

- **Qu'est-ce qu'un plasma ?**
  - **Le quatrième état de la matière :**
    - Gaz ionisé formé de particules neutres, ions et électrons libres,
    - Environnement terrestre (altitude > 80 km) sous forme de plasma,
    - Différentes régions : ionosphère, magnétosphère



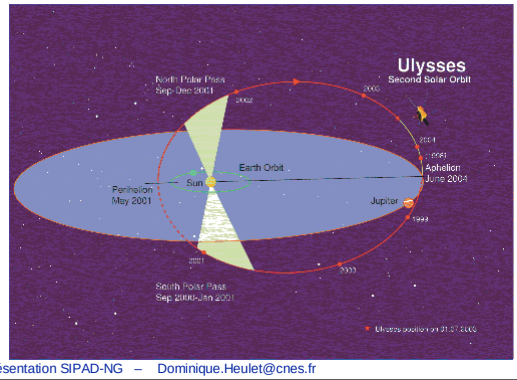
- **Dans les années 90, prise de conscience :**
  - **De l'intérêt à long terme des observations scientifiques,**
  - **Du risque élevé de perte de ces données,**
  - **➔ De la nécessité d'une entité responsable de la pérennisation des données des expériences à participation française**

Quelques missions archivées au CDPP (1/4)

- **Missions anciennes :**
  - **GEOS**
    - Mission européenne,
    - Avril 1977 ➔ décembre 1983,
    - Étude de la magnétosphère
  - **ARCAD 3**
    - Mission franco-soviétique,
    - Septembre 1981 ➔ décembre 1986,
    - Étude de la magnétosphère
  - **ISEE 3**
    - Mission NASA / ESA,
    - Août 1978 ➔ janvier 1987,
    - Satellite au Point de Lagrange : étude du vent solaire

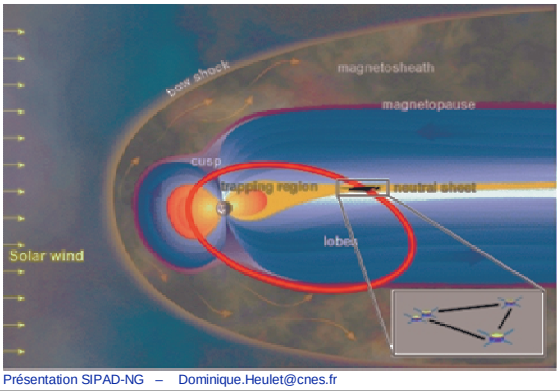
Quelques missions archivées au CDPP (2/4)

- **Missions en cours :**
  - **ULYSSE**
    - Mission NASA / ESA,
    - Depuis novembre 1990,
    - Observation des pôles du Soleil (1 orbite en 6 ans et 3 mois ➔ troisième orbite en cours)



Quelques missions archivées au CDPP (3/4)

- **CLUSTER**
  - Mission ESA,
  - Depuis février 2001,
  - 4 satellites pour l'étude des frontières entre les différentes régions de la magnétosphère,
  - ➔ 11 instruments sur chaque satellite dont 3 sous responsabilité de laboratoires français (le CDPP est donc responsable de la pérennisation des données de ces 3 instruments)



Quelques missions archivées au CDPP (4/4)

- **STEREO**
  - Mission NASA,
  - Depuis octobre 2006,
  - 2 satellites

- **Missions au sol**
  - **EISCAT**
    - Coopération Finlande, France, Allemagne, Japon, Norvège, Suède, Royaume-Uni,
    - Radars installés en Scandinavie

Caractéristiques des données archivées

- **Archivage de :**
  - **Données numériques (nombres) produites par des logiciels spécifiques :**
    - Mesures de champs (magnétique / électrique) et de particules (nombre, énergie),
    - ➔ Mesures continues
  - **Quicklooks : imagettes facilitant la sélection de données**
  - **Documents et logiciels**
  - **➔ Données le plus souvent publiques (pas de gestion complexe de droits)**
- **Formats de données :**
  - **Formats plus ou moins normalisés exploitables par des outils génériques :**
    - CDF, NetCDF, CEF, format CDPP,
    - ➔ tous ces formats présentent des limites
  - **Beaucoup de formats spécifiques des missions et des expériences**
- **Formats de métadonnées :**
  - **Travaux de normalisation en cours mais longs à aboutir,**
  - **Évolution des métadonnées ➔ nécessité de migrations régulières**

Organisation du CDPP

- **Définie par une convention CNES - CNRS (durée : 4 ans, renouvelable) :**
  - **Comité Directeur,**
  - **Comité des Utilisateurs,**
  - **Responsable Scientifique,**
  - **Composante d'Activités Scientifiques au sein d'un laboratoire toulousain :**
    - Définition des besoins d'archivage
    - Expertise sur les métadonnées
    - Développement de Services à Valeur Ajoutée (outils d'analyse propres à la thématique)
    - Site Web du CDPP : http://cdpp.cesr.fr
  - **Composante d'Activités Techniques au sein du Centre Spatial de Toulouse :**
    - Collecte et archivage des données
    - Site Web d'accès à l'archive (SIPAD-NG) : http://cdpp2.cnes.fr/cdpp/
    - Développements industriels (logiciels d'archivage, Services à Valeur Ajoutée)
- **5 scientifiques + 3 ingénieurs CNES + support industriel (3 personnes)**
  - **➔ Diversité et complémentarité des compétences nécessaires**

Quelques chiffres

- **Volume de l'archive :**
  - **10 missions,**
  - **280 jeux de données,**
  - **4 Téra-Octets**
- **Fréquence et modalités des versements très variables :**
  - **Chaînes automatiques quotidiennes :**
    - Transfert réseau laboratoire fournisseur ➔ CNES,
    - Archivage STAF,
    - Production des métadonnées et ingestion par le SIPAD-NG
  - **Chaînes mensuelles, trimestrielles**
  - **Transferts massifs par médias :**
    - Utilisation de disques externes de 1 Téra-Octets
  - **Versements ponctuels**

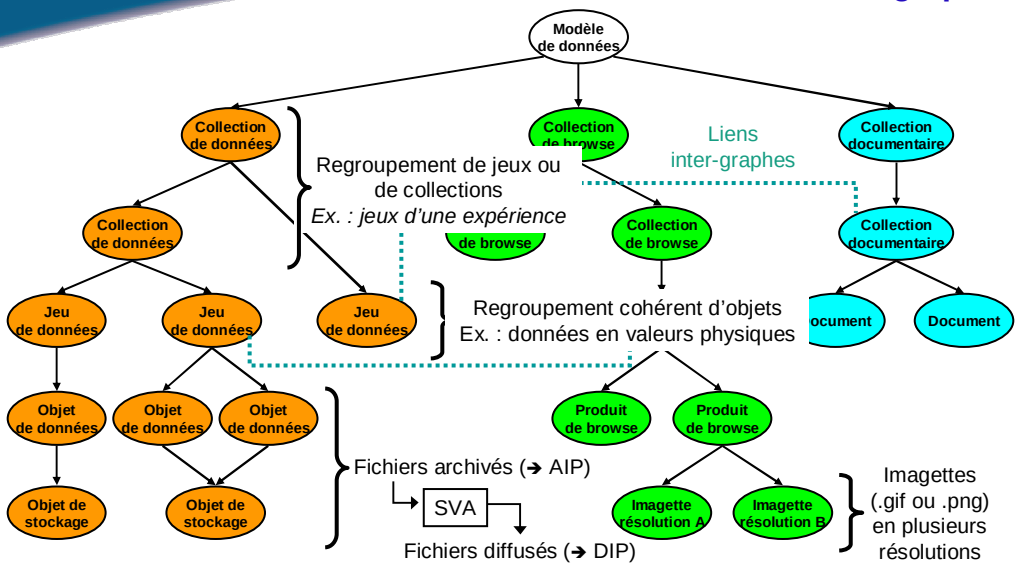
Le CDPP et le modèle OAIS

- **Démarrage du CDPP antérieur au modèle OAIS**
- **Mais on retrouve dans le CDPP de nombreux principes de l'OAIS :**
  - **Séparation des fonctions de :**
    - Stockage (service STAF),
    - Versement (modalités propres à chaque mission/expérience),
    - Gestion des données descriptives et accès aux données (SIPAD-NG)
  - **Interface Producteur - Archive clairement définie,**
  - **Importance des données descriptives**
- **➔ L'expérience du CDPP a constitué une entrée des travaux du CCSDS ayant abouti à la définition du modèle OAIS**
- **Élaboration au fil des ans d'une méthodologie d'archivage :**
  - **Documentée au travers d'un guide méthodologique,**
  - **Servant de base à la définition des procédures d'archivage des données spatiales d'autres thématiques**

Deux exemples de systèmes d'accès basés sur le SIPAD-NG

- **Serveur d'accès à l'archive du CDPP :**
  - **http://cdpp2.cnes.fr/cdpp/**
- **Serveur d'accès à l'archive des produits « vitrine » MERCATOR :**
  - **Serveur Intranet (ouverture sur l'Internet prévue en avril 2008),**
  - **Thématique : océanographie opérationnelle**
- **➔95 % de code commun entre ces deux systèmes :**
  - **Différences au niveau des IHM,**
  - **Services à Valeur Ajoutée spécifiques à la thématique :**
    - SVA d'extraction temporelle pour le CDPP,
    - SVA d'extraction de zone pour MERCATOR
- **Démonstration de ces deux serveurs**

Les graphes



Les attributs

- **Les éléments du graphe sont décrits par des attributs**
- **Attributs génériques :**
  - **Par exemple :**
    - Taille d'un objet de données,
    - Nombre d'objets dans un jeu de données
- **Attributs spécifiques à la discipline**
  - **Exemples d'attributs propres à MERCATOR :**
    - Localisation géographique d'un produit (latitudes et longitudes min. et max.),
    - Profondeur
    - ➔ Ces attributs n'auraient aucun sens en Physique des Plasmas
- **Utilisation des attributs :**
  - **Présentation d'informations associées aux données,**
  - **Sélection de données à l'aide de critères de sélection**

Exemple de déclaration d'un objet de données

```
<?xml version="1.0" encoding="UTF-8" ?>
<SIPAD_DATA xmlns="http://cnes.fr/dico_SIPAD"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://cnes.fr/dico_SIPAD SIPAD_Dictionary_DEMO_V1.0.xsd"
PROJECT_NAME="SIPDEMO">

<DATA_OBJECT_DESCRIPTION_MERCATOR_SIMPLIFIE ENTITY_TYPE="DATA_OBJECT_DESCRIPTION">
        <DATA_OBJECT_IDENTIFIER>exemple_fichier_Mercator.nc</DATA_OBJECT_IDENTIFIER>
        <ASCENDING_NODE>DA_TC_PSY1U2R1</ASCENDING_NODE>
        <RUN_DATE>2004-11-03</RUN_DATE>
        <TIME_PERIOD>
                <START_DATE>2004-10-20T00:00:00</START_DATE>
                <STOP_DATE>2004-10-20T23:59:59</STOP_DATE>
        </TIME_PERIOD>
        <GEO_COORDINATES>
                <LONGITUDE_MIN>-98.5</LONGITUDE_MIN>
                <LONGITUDE_MAX>20.0</LONGITUDE_MAX>
                <LATITUDE_MIN>-20.0</LATITUDE_MIN>
                <LATITUDE_MAX>70.0</LATITUDE_MAX>
        </GEO_COORDINATES>
        <FILE_SIZE>70</FILE_SIZE>
        <DATA_STORAGE_OBJECT_IDENTIFIER>exemple_fichier_Mercator.nc</DATA_STORAGE_OBJECT_IDENTIFIER>
</DATA_OBJECT_DESCRIPTION_MERCATOR_SIMPLIFIE>

</SIPAD_DATA>
```

Dictionnaire

Déclaration d'un objet de données

Nom de l'objet

Nom du jeu d'appartenance

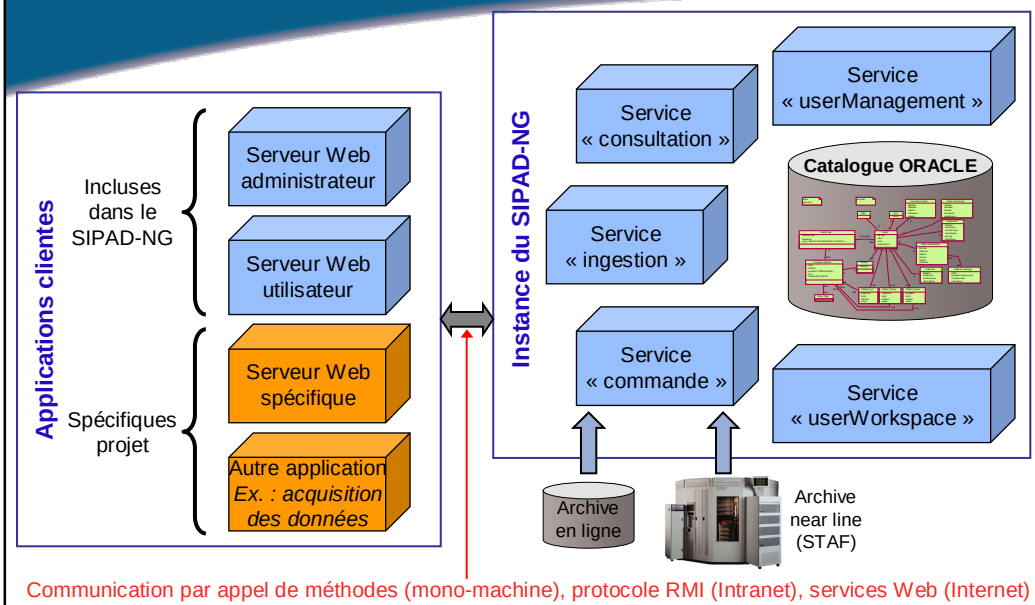
Attributs spécifiques

Attribut générique

Objet de stockage associé

Démonstration

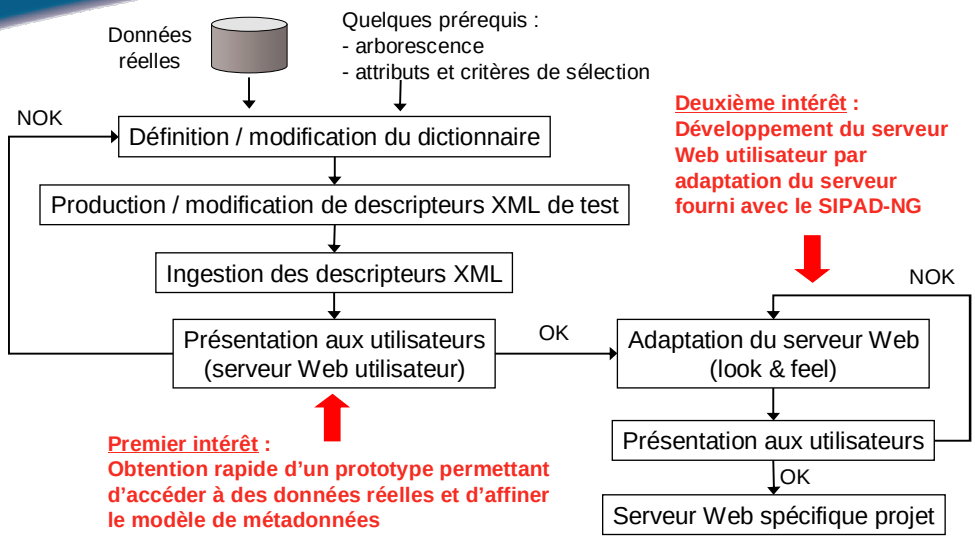
- **IHM administrateur ➔ ingestion des descripteurs XML :**
  - **Mini-arbre de données du CDPP :**
    - 2 sous-collections « mission ULYSSE » et « expérience ULYSSE / URAP »,
    - 2 jeux de données de l'expérience URAP
  - **Mini-arbre de données du modèle MERCATOR « PSY1 » :**
    - 1 sous-collection « modèle PSY1 »,
    - 1 jeu de données sur l'Atlantique Nord
  - **Mini-catalogue d'une bibliothèque :**
    - 1 jeu de documents relatifs au modèle OAIS
    - ➔ documents considérés comme des données (possibilité de recherche par critères)
- **IHM administrateur ➔ positionnement des { droits d'accès ; critères de sélection }**
- **IHM utilisateur ➔ { visualisation des pages générées ; sélection et commande de données }**



Apports de l'architecture

- **Performances tout à fait satisfaisantes :**
  - **Technologies Java (langage et outils Apache) et Oracle (procédures stockées)**
  - **➔ Nécessité de compromis généricité vs. performances**
- **Grande modularité du système :**
  - **Indépendance des composants :**
    - Utilisation optionnelle et remplacement possible (ex. : les IHM),
    - Fiabilité : l'arrêt d'un composant n'entraîne pas l'arrêt du système
  - **Déploiement possible sur tous les types d'architectures :**
    - Mono-machine,
    - Multi-machines,
    - Duplication possible des composants
- **Évolutivité :**
  - **Ajout d'applications clientes :**
    - Acquisition et archivage de données (traitement des SIP),
    - Interface avec d'autres systèmes de diffusion

Méthode de travail itérative



Enseignements et perspectives

- **Enseignements :**
  - **La migration du SIPAD vers le SIPAD-NG s'accompagne :**
    - D'un enrichissement et d'une migration des données descriptives,
    - D'une refonte complète des serveurs Web d'accès aux données
  - **Les choix de technologies et d'architecture s'avèrent pertinents**
  - **La technologie de dictionnaire de métadonnées permet :**
    - De valider les données descriptives livrées à l'archive (conformité à un schéma XML),
    - D'archiver les données descriptives avec les données (AIP),
    - De faciliter la migration vers un nouveau système d'accès aux données
- **Perspectives :**
  - **Le SIPAD-NG est utilisé par plusieurs projets :**
    - Le CDPP (Physique des Plasmas) ➔ accès à l'archive (opérationnel)
    - MERCATOR (Océanographie) ➔ accès à l'archive (opérationnel fin premier trimestre 2008)
    - SSALTO (Altimétrie et Localisation précise) ➔ développement d'une archive active
    - MINOS/SADIC (données techniques lanceurs) ➔ opérationnel (Intranet uniquement)
    - SMOS (mesure de salinité) ➔ implémentation à l'IFREMER (Brest)
    - ICARE (nuages et aérosols) ➔ Implémentation à l'Université de Lille