Un modèle d'organisation

Pour les Archives numériques

Présentation Groupe PIN

1 septembre 2004

Claude HUC (CNES)

Plan

- Notre contexte : L'archivage long terme des données spatiales
  - une approche pragmatique
- La préservation numérique : un problème commun à tous
- Le Modèle d'organisation proposé
  - la vue du Modèle abstrait OAIS
  - vue globale
  - description de chaque service : fonctions, responsabilités, interfaces externes, compétences nécessaires
- Conclusions

Le retour d'expérience au CNES

- Données spatiales sous forme numériques depuis 40 ans
- Accélération de l 'obsolescence des technologies à partir de 1990
- Programme de sauvegarde conduit de 1995 à 2000
  - motivé par la disparition annoncée des technologies de stockage sur bande magnétique
  - l 'essentiel a été sauvé mais des observations scientifiques utiles ont néanmoins été perdues
  - ==> supports détériorés mais le plus souvent une description de l information incomplète, inexacte, voire non disponible,

Le retour d'expérience au CNES

- Des documents textuels volumineux saisis sous traitement de texte en 1985
  - *saisis à nouveau sous MS Word en 1990 (Word 2) puis saisis à nouveau sous MS Word en 1997 (Word 95)*
    chaîne de compatibilité rompue en moins de 10 ans

Le retour d'expérience au CNES

- Élaboration progressive de solutions pragmatiques aux problèmes posés
- Constats d'un grand nombre de fonction à assurer
- Constat d'une grande diversité de nouveaux domaines de compétences indispensables, complémentaires des compétences nécessaires pour la préservation des documents non numériques
  - Supports et préservation des fichiers
  - Formats de représentation de l'information
  - Connaissance et compréhension de l'information numérique archivée
  - Technique de mise à disposition de cette information
- Une relative paralysie par rapport à l'évolution très rapide des technologies
  - Coûts induits
  - Comment s'y prendre

Autres constats

- Une focalisation excessive sur les questions techniques (choix de supports, formats)
- Une distinction insuffisante entre :
  - des cas que l'on sait résoudre et pour lesquels on peut proposer des solutions techniques et organisationnelles viables
  - des cas que l'on ne sait pas (pas du tout ou pas de façon satisfaisante) et pour lesquels les approches proposées relèvent encore du domaine de la recherche et de l'expérimentation (émulation)
- Des solutions dangereuses à long terme (migrations continues des formats)
  - qui sont encore présentées comme viables alors que ce ne sont que des palliatifs à court terme

Le problème : la réduction des échelles de temps

- Problèmes rencontrés :
  - *documents bureautiques :* ==> **chaîne de compatibilité rompue en moins de 10 ans**
  - Des documents scientifiques plus récents (1995 !!) mais pour lesquels il a fallu ressaisir l'ensemble des formules mathématiques.
  - Des observations scientifiques enregistrées sur bandes magnétiques sauvées in extremis
- Accélération des évolutions de la technologie depuis les années 90
  - ce mouvement ne fléchit pas, bien au contraire (au moins 5 versions de MS Word sous Windows de 1995 à aujourd'hui).

Un document numérique enregistré il y a 10 ans ou parfois moins peut déjà être dans une situation vulnérable au regard de sa préservation.

Qui est aujourd'hui concerné ?

- Pratiquement tout le monde :

  - administration : état civil…
  - secteur de la santé,
  - caisses de retraite,
  - l'industrie : pétrole, aéronautique...
  - la recherche scientifique, le domaine spatial
  - la Défense,
  - le nucléaire,
  - …..et aussi les particuliers.

Vulnérabilité : des causes multiples

- Facteurs techniques
  - obsolescence des technologies de stockage, des logiciels et des systèmes
  - dépendances entre données créées et l'environnement de création
  - données ou documents non décrits

- **Facteurs organisationnels et financiers**
  - **la pérennisation de l'information constitue une activité en soi.**
  - **l'organisation du travail, le partage des responsabilité, la mise en place des bonnes compétence au bon endroit sont à repenser**
- Facteurs normatifs, juridiques, industriels, psychologiques, liés à l 'absence de formation fondamentale...

Comprendre le problème posé pour le résoudre

- C'est l 'objet du ' Reference Model for an Open Archival Information System (OAIS) ' Issue 1. January 2002

http://www.ccsds.org/CCSDS/recommandreports.jsp#interchange
Référence CCSDS : CCSDS 650.0-B-1 (gratuit)

http://www.iso.org/ Référence ISO : ISO 14721:2002 (206 francs suisses…)

Traduction en cours en collaboration entre le CNES et la BnF

- Analyse détaillée, définition des concepts, d 'un modèle fonctionnel et d 'un modèle d 'information pour comprendre l 'ensemble des spécificités de l 'archivage de l 'information sous forme numérique

OAIS :Qu'est ce qu'une archive ?

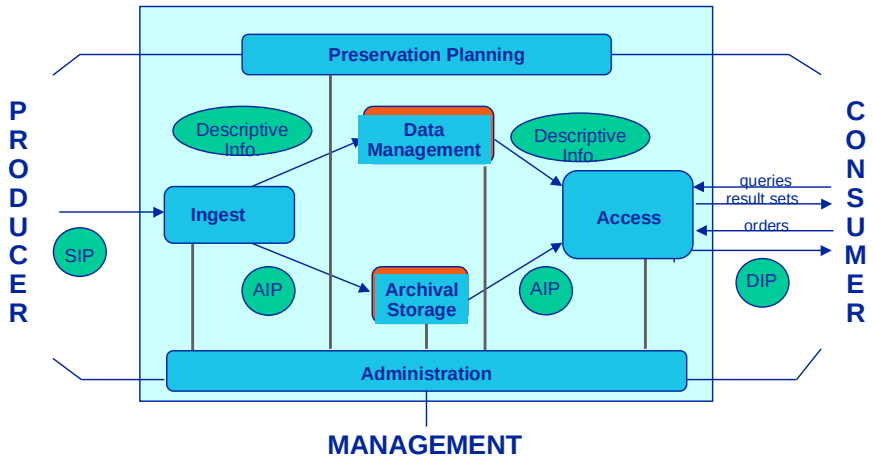
- Organisation dont la vocation est de préserver l'information pour permettre à une **Communauté Définie d'Utilisateurs** d'y accéder et de l'utiliser.
  - pérennité des données
  - accès pérenne aux données
  - préservation avec les données, de toutes les informations nécessaires à leur compréhension et utilisation

  Définition extraite de la norme ISO 14721:2002
- I 'archivage n'est
  - ni une sauvegarde, ni un backup système
  - ni un rangement définitif des données quand on pense qu'elles ne serviront plus

OAIS : Données et information

- **Information**
  - tout type de connaissance pouvant être échangée
  - indépendante des formes (à savoir, physique ou numérique) utilisées pour représenter cette information
- **Données** : formes de représentation de l'information

OAIS : entités fonctionnelles

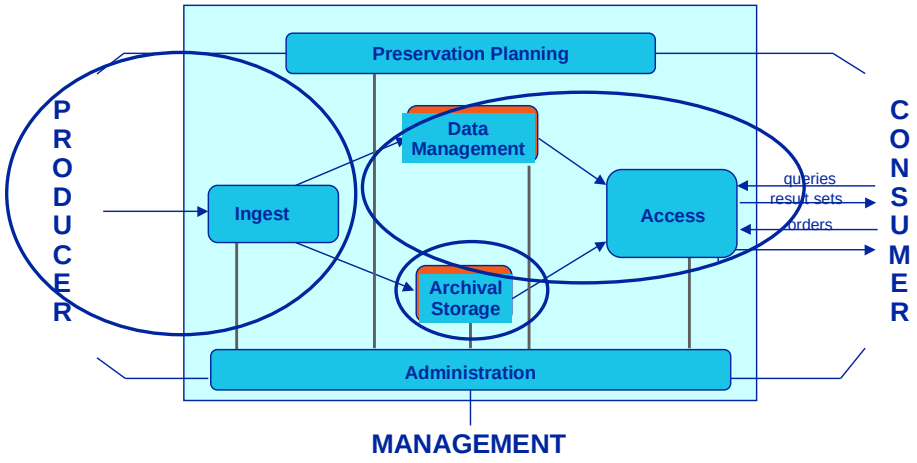


SIP = Submission Information Package
AIP = Archival Information Package
DIP = Dissemination Information Package

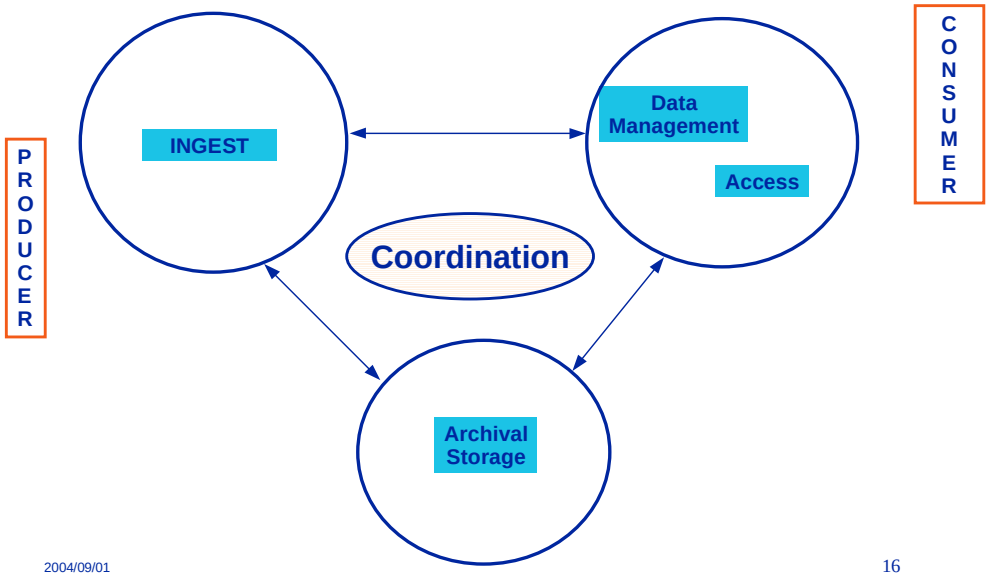
OAIS : entités fonctionnelles

- Un très grand nombre de fonctions élémentaires à prendre en charge
- On ne sait pas bien par quel bout prendre le problème
- L'organisation en place pour les archives de documents non numériques n'est pas forcément adéquate
- D'où l'idée de proposée une organisation en services indépendants chargé de fonctions précises et dotés d'interfaces parfaitement définies : réduire la taille du problème pour pouvoir le résoudre
  - faire le lien entre l'approche abstraite globale de l'OAIS
  - et les solutions pragmatiques expérimentées

Recherche d'une organisation pratique



Trois services coordonnés

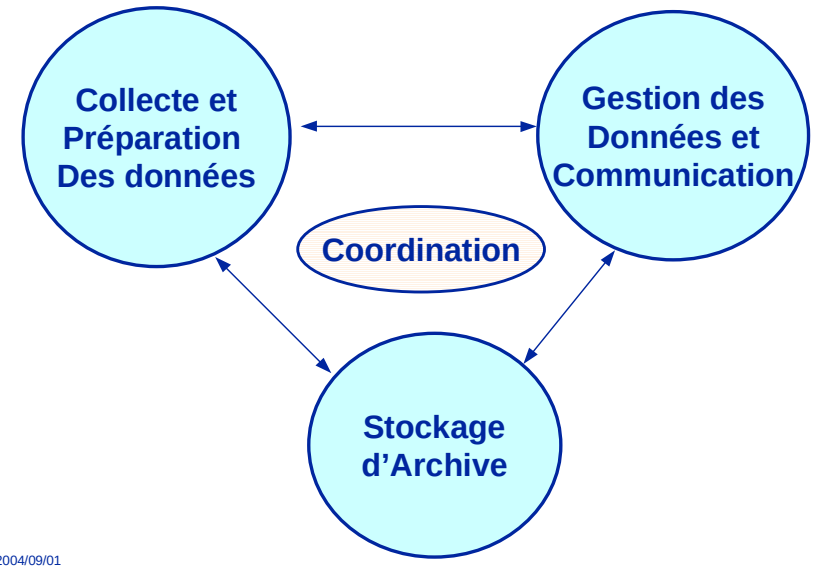


Service ?

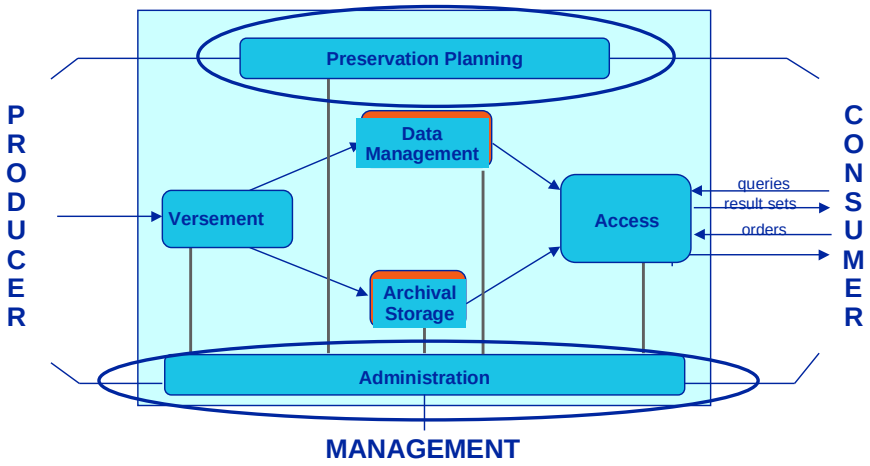
- **Le terme 'Service' est défini ici comme un ensemble composé de personnes, de moyens techniques et de ressources financières ou autres, en charge d'un mandat clairement identifié**
- **Il s'agit de montrer que pour chaque service proposé, on peut :**
  - définir précisément les fonctions les responsabilités,
  - spécifier les interfaces externes (relations avec les autres services et relations avec les entités externes à l'archive),
  - de préciser les compétences nécessaires à son fonctionnement.

Chaque service est supposé devoir mettre en application un ensemble de procédures et de normes qui lui sont propres et disposer des moyens et ressources adaptés aux tâches dont il a la charge.

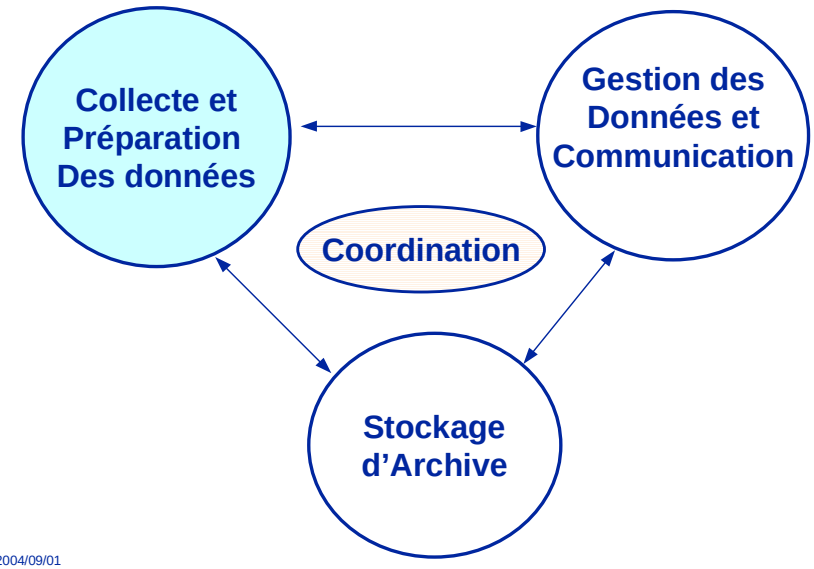
Trois services coordonnés



Considérations complémentaires



Le service 'Collecte et préparation des Données' (CPD)



Le service 'Collecte et préparation des Données' (CPD)

Responsable de :

- La collecte des objets numériques auprès des producteurs
- L'ensemble des tâches permettant d'ajouter à ces objets les informations nécessaires à leur préservation
- L'ensemble des tâches permettant de passer des objets livrés par le producteur à des objets ayant les qualités requises pour être préservés : changement de formats éventuel s'ils sont autorisés
- Le service CPD prend en charge :
  - l'ensemble des actions et tâches identifiées dans le standard CCSDS et projet de norme ISO ' Producer-Archive interface abstract standard '
    - négociation : ce que le producteur peut et ne peut pas faire
  - les transformations sur les données et métadonnées restant à la charge du service d'archive dans cette négociation

Service CPD : principales tâches (1/2)

- assurer la réception des objets transmis par les services versants et contrôler leur conformité par rapport au plan établi,
- effectuer lorsque cela est nécessaire, des opérations de transformation de format de données et de métadonnées
  - (par exemple des fichiers livrés au format MS Word pourront être transformés en fichiers au format PDF/Archive, des fichiers texte contenant des métadonnées pourront être transformés en fichiers XML structuré),
- affecter aux objets numériques reçus, un identifiant unique cohérent dans l'espace de nomenclature de l'Archive,

Service CPD : principales tâches (2/2)

- enrichir les métadonnées en mettant les objets reçus en relation contextuelle avec d'autres objets déjà archivés, ou avec des documents disponibles dans d'autres Archives,
- transférer tous objets numériques archivables (données et métadonnées) au service de Stockage d'Archive,
- transférer les métadonnées et éventuellement des objets ayant vocation à être disponibles en ligne au service Gestion des Données et Communication.

service CPD : interface externe

➢ interface avec le service 'Archival Storage' (AS)

- les fichiers de données et de métadonnées à préserver ont été transmis au service 'Archival Storage' au sein duquel, ils ont été organisés dans une arborescence 'virtuelle'
- le service Archival Storage' est en charge de leur préservation

➢ interfaces sont extrêmement simples. Elles se résument à un tout petit nombre d'actions qui peuvent être mises en œuvre depuis un poste de travail du service CPD :

Voici une représentation réaliste des actions permettant au de transmettre un objet numérique au service AS :

- *Connexion au service AS (toujours sur l'initiative du Service CPD) et authentification*
- *Demande de prise en charge du stockage d'un objet numérique pour lequel on précise : son identifiant et la classe de service attendue pour cet objet*
- *Transmission du fichier*
- *Accusé de réception du service AS*
- *Fermeture de session*

service CPD : interface externe

- En sortie : interface avec le service ‘Gestion des Données et Communication’
  - les fichiers de métadonnées, sous une forme normalisée, sont transmis au service GDC
  - ces fichiers de métadonnées sont de tous niveaux : ils peuvent inclure
    - des descriptions de collections et de sous-collections (fonds et sous-fonds),
    - mais aussi des descriptions et des identifications d'objets numériques unitaires.
    - Le service CPD peut également transmettre dans ce cadre, des objets numériques spécifiques qui être utile à la recherche d'information dans l'Archive (par exemple des représentations graphiques).

service CPD : les choix essentiels

- Nous avons retenu les choix essentiels suivants :
  - les données numériques (documents issus de la bureautique, observations scientifiques, images, vidéo…) doivent être sous une forme (un format)
    - de préférence normalisé
    - indépendant des logiciels mis en œuvre pour les créer
    - décrites (syntaxe et sémantique) de façon exhaustive
  - les métadonnées doivent être normalisées
- Nous avons rejeté les voies s ’appuyant sur des migrations de format régulières et incertaines des données

service CPD : moyens techniques nécessaires

- Les moyens matériels, logiciels et de communication nécessaires à la bonne réception des objets numériques transmis par les services versants ne présentent pas de caractéristiques spécifiques,
  - nécessité à décider au cas par cas, de sécuriser les transferts
    - afin d'authentifier les objets reçus
    - et de garantir leur intégrité par rapport à l'expéditeur.
- Ces moyens seront à adapter en fonction du volume des données à prendre en compte et de la périodicité des transferts.
- Un ensemble de logiciels d'aide à la préparation des données et des métadonnées à archiver sera naturellement nécessaire.

service CPD : compétences requises

- Il apparaît clairement ici un besoin de double compétence :
  - celle de l'archiviste capable
    - de définir, en relation avec le producteur, les informations à préserver
    - de vérifier l'intelligibilité de ces informations et leur complétude
    - d'organiser ces informations au sein d'un ensemble structuré,
  - celle de l 'informaticien spécialisé dans la **gestion des données** et la représentation de l 'information sous forme numérique, afin
    - de définir les formats de données et de métadonnées acceptables pour la pérennisation
    - de vérifier la conformité
    - de mettre en œuvre si nécessaire un processus de transformation de formats
    - de spécifier le développement des outils informatiques nécessaires à ce service, de les développer et de les exploiter

    Ces compétences spécialisées sur la représentation numérique présupposent également une connaissance généraliste en Informatique.

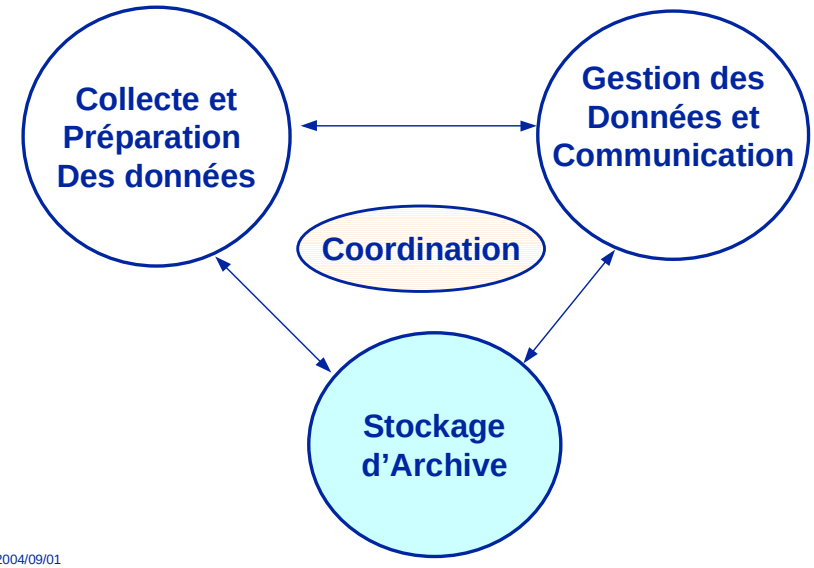
service CPD : compétences requises

- Les deux compétences sont réunies dans un métier nouveau qualifié de gestionnaire de données numériques s'appuyant très fortement sur les normes et standards de représentation de données et de métadonnées
- En outre, il apparaît la nécessité de pouvoir dialoguer, négocier avec les entités productrices de données et de documents :
  - travail itératif de longue haleine

service CPD : Retour d'expérience au CNES

La collecte d'un ensemble complet, organisé et convenablement décrit d'objets numériques constitue dans la pratique la tâche la plus difficile et en définitive la plus coûteuse, en particulier en ressources humaines.

STOCKAGE d'ARCHIVE



Stockage d'Archive : Fonctions vue du coté ' client '

- Scénario :
  - je suis responsable de l'archivage de données digitales, images, documents, vidéo…

  Ces données sont des ensembles de fichiers
    - c'est à dire des trains de bits
    - dont je connais le format et le contenu
    - que je sais manipuler et présenter aux utilisateurs finaux sous une forme intelligible
  - ce que j'attend d 'un service de Stockage d'Archive, c'est d'abord :
    - la prise en charge de ces fichiers en vue de leur conservation à long terme
    - la garantie de l'intégrité de ces fichiers
    - la capacité à me restituer ces fichiers dans le délai convenu par le contrat de service
    - la disponibilité d 'une ' **interface technique ' stable** me permettant de faire appel à ses services (archivage de fichier, restitution, renommage, création d 'arborescence virtuelle,…)
    - la capacité à prendre en compte les évolutions de la technologie (migrations de supports de stockage….) sans aucun impact sur l 'interface et donc sans impact sur mes applications
    - La gestion des droits d'accès à ces données
  - ce concept permet :
    - une organisation du service de Stockage d'Archivetotalement indépendante des autres services
    - une réutilisation de ce service dans de multiples contextes au sein de l 'organisme concerné

Stockage d'Archive (SA) : responsabilité de l'intégrité des fichiers

- Le SA doit prendre en charge l'ensemble des activités nécessaires au maintien de l'intégrité des objets numériques :
  - stockage des objets sur des médias de stockage, accompagné d'une ou plusieurs copies de sauvegarde devant être entreposées dans des locaux séparés,
  - surveillance permanente de l'état des médias (nombre d'opérations de lecture réalisée sur chaque média, taux d'erreur de bits mesurable…),
  - remplacement périodique des médias jugés moins fiables par des médias neufs,
  - prise en compte des évolutions des technologies de stockage pour opérer des migrations (périodiques ou continues suivant la politique retenue) vers les nouveaux médias les plus appropriés à ses activités.
  - Etc.

Stockage d'Archive : Compétences

- Compétences d'informaticiens spécialisés dans :
  - la gestion de grands ensembles de fichiers stockés, dupliqués sur différents types de supports,
  - les technologies réseau à haut débit permettant de communiquer avec les ' clients ' du service
  - les technologies de stockage à grande capacité, robots de stockage…, les supports de stockage, leurs caractéristiques, leur fiabilité
  - les moyens de surveillance de l 'état des supports, mise en œuvre de ces moyens
  - la capacité à maintenir en fonctionnement opérationnel un système ouvert 24 heures sur 24 et à faire évoluer le système en fonction des évolutions de la technologie et des montées en charge

Stockage d'Archive : le service mis STAF en place au CNES

- Ce service est le STAF (Service de Transfert et d'Archivage de Fichiers)
- mis en place en 1994

  - le Service STAF a pour mission de pérenniser les données patrimoniales du CNES issues d'expériences scientifiques.
  - ce sont des données de référence, non reproductibles, stables dans le temps et destinées à être utilisées sur le long terme
  - l'idée du STAF : permettre de ranger une « collection » de fichiers selon une logique applicative et sans se soucier des évolutions des systèmes et des technologies

Stockage d'Archive : le service mis STAF en place au CNES

- Garantie de l'intégrité et de la confidentialité des données de chaque 'client' utilisateur du service
- Transparence des opérations d'exploitation
- Possibilité d'étendre au fur et à mesure les capacités de stockage
- Capacité de prise en charge de nouvelles machines clientes
- Actuellement : plus de 3,8 millions de fichiers pour un volume de 145 Terabytes

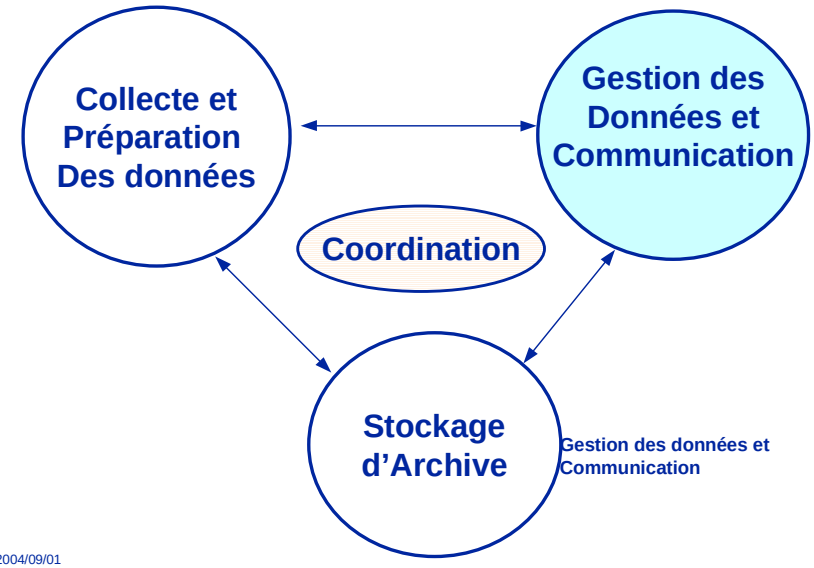
Retour d'expérience au CNES

- Le concept du STAF - Service d 'Archivage de Fichiers (stockage d 'archive) a totalement fait ses preuves
  - dix ans d 'expérience
  - aucune donnée perdue
  - un nombre croissant de clients
  - un volume croissant de données stockées
  - le concept permet aussi la mutualisation du stockage d 'archive
    - entre plusieurs établissements d 'une même institution
    - entre plusieurs institutions distinctes
  - **une masse critique minimale est indispensable pour réduire les coûts (des moyens matériels, logiciels et humains)**

Retour d'expérience au CNES

- Le principe de fonctionnement de ce service et les résultats pratiques obtenus depuis 10 ans ont convaincus la BnF de l'intérêt d'un tel service
- Un projet de convention entre le CNES et la BnF, portant sur la réutilisation, par le BnF, des logiciels de gestion du service, est actuellement envisagé
- Un tel service peut :
  - être propre à une institution
  - être partagé par plusieurs institutions distinctes
  - être pris en charge par une entreprise privée

Gestion des données et Communication



Gestion des Données et Communication (GDC): fonctions

- Responsable de la gestion du patrimoine d'information préservé par l'Archive et de la Communication de ce patrimoine auprès des utilisateurs autorisés.
- Mise en place et maintien en fonctionnement d'un système informatique permettant aux utilisateurs d'accéder à distance - via une interface graphique - à un ensemble de fonctions
  - connaître le contenu de l 'archive,
  - rechercher les données qui les intéresse (critères de sélection basés sur les métadonnées par exemple)
  - sélectionner les données qui correspondent à leur besoin
  - commander et récupérer ces données
  - éventuellement transformer les données archivées avant de les fournir à l 'utilisateur (changements de format, Services à Valeur Ajoutée….)

Gestion des Données et Communication: fonctions

- La recherche des données utiles s 'appuie sur les métadonnées mais aussi sur différentes techniques complémentaires (feuilletage, data mining…)
- Les moyens de récupération peuvent être le réseau ou la recopie sur un support de diffusion courant (CD-rom, DVD, DLT…) en fonction du volume
- Gérer les relations avec la communauté des utilisateurs

Gestion des Données et Communication: moyens techniques nécessaires

- **Moyens techniques nécessaires**
  - le système mis en place par le service GDC s'appuie largement sur les technologies de base de données et de communication d'information via Internet.
  - des systèmes répondant partiellement ou totalement aux besoins du service GDC sont ou seront disponibles sur le marché, ce qui limitera le coût des développements informatiques spécifiques.
  - GDC doit éventuellement disposer de capacité de recopies d'objets numériques sur les médias de diffusion. Enfin, dans certain cas, indépendamment du service SA, il peut être amené à stocker à son niveau les objets de données qui ont vocation à être immédiatement disponibles en ligne pour les utilisateurs, d'où un besoin d'une capacité de stockage (généralement sur disque) à cet effet.

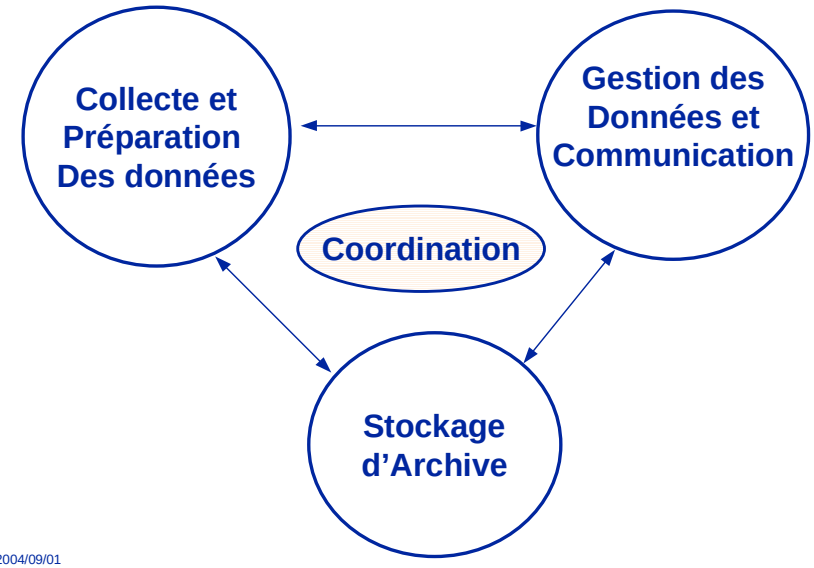
Gestion des Données et Communication: compétences requises

- Compétences d 'informaticien spécialisé dans :
  - la modélisation des données
  - les processus de recherche d 'information
  - les technologies de base de données
  - les technologies et langages de l 'Internet (Interface Homme-machine sur navigateur,….)
  - le maintien en fonctionnement opérationnel de systèmes ouverts à des communautés d'utilisateurs plus ou moins vastes
- Connaissance générale de la problématique de l 'archivage
  - connaissance des catégories de données manipulées
  - connaissance des métadonnées et des critères de sélection des données adaptés aux besoins des utilisateurs

Gestion des Données et Communication : retour d'expérience au CNES

- De tels capacités de Gestion de Données et de Communication ont été mises en œuvre
  - dans le but de mettre à disposition des données scientifiques spatiales de différentes thématiques (astronomie, océanographie…).
  - malgré la diversité des objets numériques et des logiques propres à chacune de ces thématiques, le défi, en passe d'être résolu, est de réduire les coûts par l'usage d'un système générique adaptable à toutes ces thématiques.

La coordination,



Le coordonnateur

- Le coordonnateur est le chef d 'orchestre et le véritable responsable de l'archivage (au sens OAIS)
  - suivant le contexte, on parlera de gestionnaire des données, de gestionnaire du patrimoine technique, d 'archiviste, d'archiviste principal,...
- Son rôle :
  - organiser le partage du travail entre les différents ' services '
  - assurer la clarté des interfaces entre ces services
  - coordonner le travail pour les domaines de compétence communs :
    - le modèle d 'information
    - le dictionnaire des objets numériques livrables
    - ....

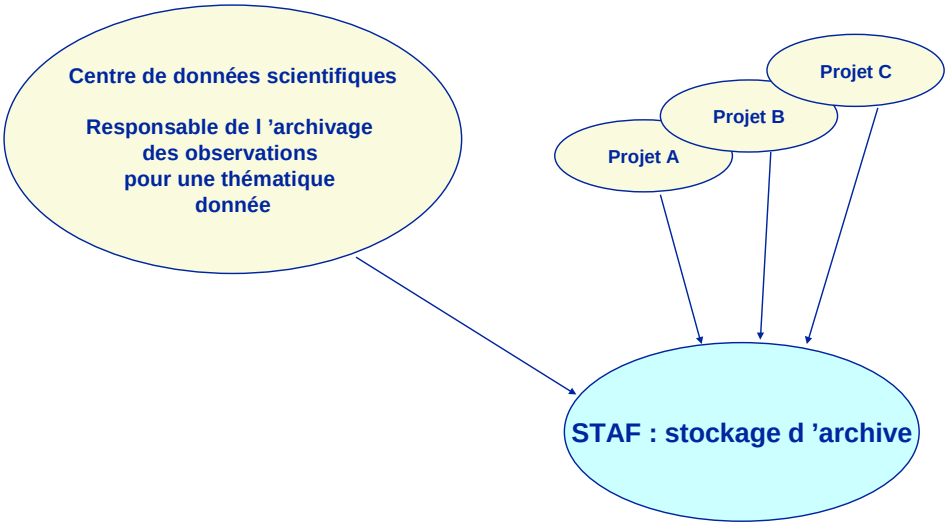
Pourquoi un Modèle d'organisation ?

- Pour identifier l'organisation adéquate des personnes et des moyens
  - à partir du Modèle, plusieurs scénarios d'organisation réels sont envisageables
- Pour préciser les compétences requises au niveau de chaque service
- L'analyse des activités et ressources de chaque service facilite l'évaluation des coûts
- Pour contribuer à la définition de produits du marché qui pourraient assurer une part significative des activités de tel ou tel service
- Pour simplifier la réflexion sur la certification des archives

Les produits du marché

- Aucun système matériel-logiciel ne peut prétendre à assurer l'éventail des fonctions de l'archivage numérique, par contre on entrevoit des possibilités :
  - au niveau du service de stockage
  - au niveau du service de gestion des données et communication
- Ainsi que des aides à la prise en charge des fonction s de collecte et préparation

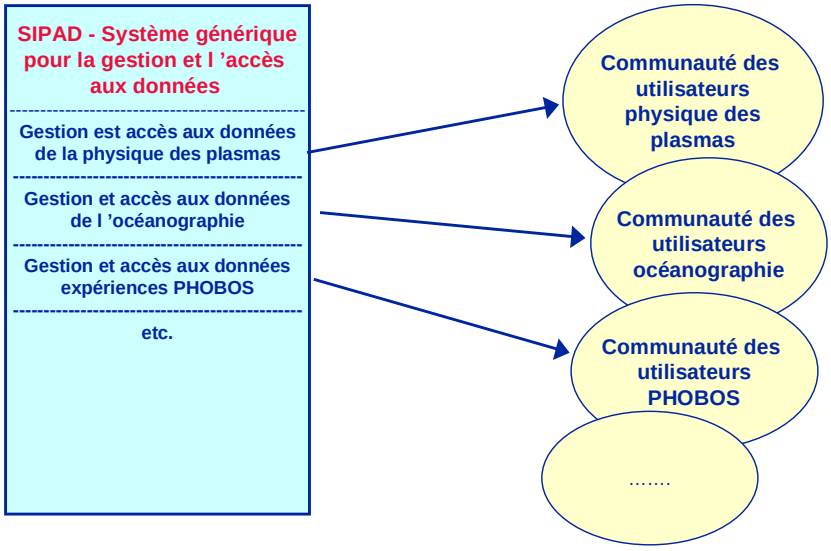
Relations avec le Système d'information de l'institution : cas du STAF



Stockage d'archive : variantes organisationnelles

- Parmi les variantes organisationnelles possibles, nous pouvons penser :
  - à un service de Stockage d'Archive partagé par plusieurs Archives distinctes,
  - à un service SA partagé entre des archives et d'autres services du même organisme (c'est le cas au CNES),
  - à un service SA sous la forme d'un prestataire de service indépendant.

Relations avec le SI de l'institution : la gestion et la mise à disposition de l 'information



Point critique essentiel

- Le point critique est et celui de la collecte de l'information et de l 'ensemble des activités conduisant à la constitution :
  - de fichiers dont le format est acceptable pour la préservation à long terme
  - de métadonnées ' normalisées '
- Ce point critique concerne à la fois :
  - le contenu (complétude, exactitude, authenticité)
  - le format (ouvert, normalisé,…)
- Ce point critique n'est pas sans rapport avec la politique technique ou la politique bureautique de l 'entreprise

Conclusion

- la technologie va évoluer sans interruption mais l'information va rester
- **C'est pour cette raison que nous avons consciemment privilégié les voies qui s'appuient sur une connaissance de la structure, de la syntaxe et de la sémantique de l'information plutôt que de tenter de maintenir une quelconque technologie en état de marche sur le long terme**
- Le Modèle d'organisation proposé repose d'abord sur ce choix.
  - sa raison d'être est de **contribuer à l'émergence de solutions concrètes et applicables**.
  - Il repose également sur une analyse des compétences et des métiers.
  - Il s'appuie enfin sur un large retour d'expérience au CNES qui nous conforte dans cette voie.

Conclusion

- Une telle organisation doit pouvoir faire l'objet de contrôles et d'audits externes.
- L'Archive numérique doit pouvoir apporter la démonstration,
  - au travers de son organisation,
  - de ses moyens,
  - de ses équipes et des standards et procédures applicables,

➔ de sa capacité à assurer sa mission et donc à préserver à long terme les informations sous forme numérique dont elle a la charge

➔ Ceci nous ouvre un champ de réflexion sur la 'Certification' des Archives Numériques.

Conclusion : notre vision pour le futur

- Le service CPD :La valeur Ajoutée intellectuelle est très forte et ne saurait être remplacée par des processus automatiques. Certains logiciels peuvent constituer une aide ai travail mais ils ne réfléchissent pas à notre place
- Le service 'Archival Storage' : fondamentalement technologique. Les industriels doivent pouvoir apporter pour ce type de service, des solutions clé en main, fiables et économiques
- Le Service 'Data management and Access' : fortement technologique mais plus dépendant du modèle d'information. Là encore, des systèmes clé en main de gestion et mise à disposition de données largement réutilisables dans différents domaines peuvent être développés