



L'archivage pérenne du document numérique au CINES

CINES (O.Rouchon)

Journée d'étude URFIST Bordeaux – 23 Mai 2008

- La mission d'archivage du CINES
- Le projet PAC
 - Les dates clé
 - Le contexte, la problématique et les constats
 - Les défis, orientations et choix pour l'archivage au CINES
 - Les normes et standards adoptés
- Les types de documents à archiver
 - La structure du document à archiver
 - Les formats de fichier et leur contrôle
- Les acteurs
- Les échanges
 - L'architecture logique de la plateforme
 - Les principes de fonctionnement
 - Les étapes des différents échanges
- L'état des lieux



UNIVERSITÉ DE
BORDEAUX



UNIVERSITÉ MONTESQUIEU
BORDEAUX IV



Centre Informatique National de l'Enseignement Supérieur

- Basé à Montpellier (Hérault, France)
- Créé en 1999, succédant au CNUSC (Centre National Universitaire Sud de Calcul) – créé en 1980
- Placé sous la tutelle de la DGRI (Direction Générale de la Recherche et de l'Innovation) et de la DGES (Direction Générale de l'Enseignement Supérieur) du Ministère de l'Enseignement Supérieur et de la Recherche
- Principales missions
 - Calcul numérique intensif
 - Archivage pérenne de documents électroniques
 - Hébergement et suivi de serveurs d'applications
- Plus d'information : <http://www.cines.fr/>



La mission d'archivage du CINES

Depuis 2004, le CINES travaille sur la mise en place d'un service pour l'archivage pérenne du patrimoine scientifique.

Cette mission a été confirmée par plusieurs décisions des organismes de tutelle :

- Arrêté du 7 août 2006 relatif aux modalités de dépôt, de signalement, de reproduction, de diffusion et de conservation des thèses ou des travaux présentés en soutenance en vue d'un doctorat
- Convention du 2 mai 2007 (faisant suite à celle du 15 octobre 2003) relative à la mise en ligne et l'archivage pérenne de données numérisées dans le cadre du programme Persée
- Lettre de cadrage du 12 février 2008 recentrant les activités du CINES autour de deux missions stratégiques :
 - le calcul intensif
 - l'archivage pérenne

Pour remplir la mission d'archivage, le CINES a mis en place le projet PAC, qui vise à se doter d'une plate-forme et d'un service d'archivage numérique pérenne

- L'équipe actuelle
 - 1 chef de projet
 - 4 ingénieurs
 - 1 archiviste
- 2 projets pilotes en cours
 - Archivage des thèses électroniques
 - Documents nativement au format électronique
 - Archivage des revues SHS du portail Persée
 - Documents issus de la numérisation de revues au format papier
- Contraintes
 - Besoin d'une solution générique, basée sur les standards pour une évolution,
 - Facilité de veille technologique et de migration logique (changement de format)

- 2004**
 - Démarrage du projet, constitution de l'équipe
 - Etude des normes, participation au groupe de travail PIN

- 2005**
 - Etude du cabinet Ourouk pour l'aide au choix d'une solution logicielle pour l'archivage électronique
 - Décision d'un développement interne, aucune solution existante n'étant satisfaisante

- 2006**
 - Documentation des spécifications fonctionnelles pour la solution souhaitée
 - Développement en interne d'une première plateforme

- 2007**
 - Premier tests d'intégration avec l'outil STAR de l'ABES pour le versement de thèses électroniques
 - Déploiement de PAC v1.0 en production et démarrage de l'exploitation

- 2008**
 - Appel d'offre pour une plateforme PAC v2.0 notifié à la société SUN
 - Déploiement prévu au deuxième trimestre

Les défis, orientations et choix pour l'archivage au CINES

L'archivage pérenne des documents électroniques consiste à conserver le document et l'information qu'il contient :

- Dans son aspect physique comme dans son aspect intellectuel,
- Sur le très long terme soit 30 ans et au-delà,
- De manière à pouvoir le rendre accessible et compréhensible.

Or, la plupart des fichiers informatiques de plus de 10 ans sont aujourd'hui illisibles, conséquences de plusieurs risques inéluctables

| Risque | Solutions |
|---|---|
| Connaissance perdue du contenu | <ul style="list-style-type: none">• Utilisation de métadonnées• Identification unique et pérenne des documents archivés |
| Format de fichier inconnu | <ul style="list-style-type: none">• Privilégier les formats durables• Identification, validation des formats• Migration logique |
| Support physique détérioré | <ul style="list-style-type: none">• Gestion du vieillissement des médias• Migration physique |
| Logiciel ou matériel de lecture disparu | <ul style="list-style-type: none">• Veille technologique et anticipation |

Les normes et standards utilisés

- OAIS - ISO 14721 : Reference model for an Open Archival Information System
 - Modèle purement conceptuel, ne fait aucune recommandation technique
- P2A Politique et pratiques d'archivage (sphère publique)
 - Recommandations en termes d'architecture, moyens, sécurité, etc.
- Normes internationales de description archivistique
 - ISAD-G – international standard for archival description, general
 - ISAAR-CPF – international standard archival authority record, corporate bodies, persons & families
- Standard d'échanges de données pour l'archivage électronique, versement, communication, élimination
 - DAF, DGME, version 1.0, mars 2006.
- Métadonnées descriptives de l'archive
 - DCMI – Dublin Core Metadata Initiative
- Identifiant unique et pérenne
 - Interne, séquentiel, basé sur le principe URI
 - Pas d'utilisation pour le moment d'identifiant persistant externe de type Handle, DOI, PURL ou ARK
- Empreintes numériques
 - Hashing MD5, SHA-256

Modèle conceptuel pour l'archivage de documents (numériques en particulier).

- Référence décrivant dans les grandes lignes les fonctions, les responsabilités et l'organisation d'un système qui voudrait préserver de l'information, en particulier des données numériques, sur le long terme.
 - Donne une terminologie fiable et unique pour manipuler tous les concepts liés à la préservation des données numériques
 - Propose les questions à se poser au moment de mettre en place un système de préservation
 - Décrit les composantes d'un tel système au niveau de l'organisation interne et externe
- Défini au départ dans le domaine aérospatial, par le CCSDS qui est l'organisme de normalisation de ce domaine.
- Aujourd'hui il est très largement adopté au-delà de cette communauté et il a été reçu comme norme par l'ISO sous le numéro 14721.

http://vds.cnes.fr/pin/documents/projet_norme_oais_version_francaise.pdf

La P2A – Politique et pratiques d'archivage

Décrit les concepts de politique d'archivage et de déclaration des pratiques d'archivage

- Politique d'archivage (PA) :
 - Exigences minimales en termes juridiques, fonctionnels, opérationnels, sécurité, etc.
 - Contraintes en matière d'authentification de l'origine de l'archive, intégrité, lisibilité, pérennité des archives, traçabilité des opérations, accessibilité, etc
- Déclaration des pratiques d'archivage (DPA) :
 - Définition des moyens pour répondre aux objectifs et engagements de la PA
- Concerne la sphère publique – archives publiques telles que définies par le Code du patrimoine
- Reprend les concepts du modèle OAIS
- Version 1.0, juillet 2006

<http://www.ssi.gouv.fr/fr/confiance/documents/methodes/ArchivageSecurise-P2A-2006-07-24.pdf>

Le Standard d'échanges DAF/DGME

Standard d'échanges de données pour l'archivage électronique, versement, communication, élimination

- Direction des archives de France, Direction générale pour la modernisation de l'Etat
- Version 1.0, mars 2006
- Cadre normatif pour
 - Les différents échanges d'information entre les services d'archives et leurs partenaires
 - La conservation des documents et données numériques
 - La spécification d'une architecture technologique de référence

http://synergies.modernisation.gouv.fr/IMG/pdf/archives_echanges_v0-1_description_standard_v1-0.pdf

Les normes internationales de description archivistique

Recommandations pour

- L'identification et la description du contexte et du contenu de documents d'archives
- Le classement et l'indexation
- ISAD-G
 - International Standard For Archival Description (General)
 - Règles générales pour la description d'archives
 - Version 2, septembre 1999
- ISAAR-CPF
 - International Standard Archival Authority Record (Corporate Bodies, Persons and Families)
 - Relatif aux archives des collectivités
 - Version 2, 2004

Conseil international des archives, ICA

<http://www.ica.org/>

Le standard de métadonnées DCMI

Dublin Core Metadata Initiative

- Standard de métadonnées interopérables
- Propose une liste de métadonnées – avec plusieurs niveaux de détails – pour qualifier les documents
- Version 1.1, décembre 2006

<http://dublincore.org/>

L'identifiant unique et pérenne

Attribution d'un identifiant unique et pérenne aux documents au moment de leur archivage (AIP) pour permettre de les référencer, et assurer une gestion cohérente de l'archive, conforme au modèle OAIS (information d'identification)

Pas d'utilisation d'identifiant persistant de type

- Systèmes de publication électronique : Handle, DOI (Digital Object Identifier) ex. : [DOI:10.1016/j.compmedimag.2006.07.004](https://doi.org/10.1016/j.compmedimag.2006.07.004),
- Systèmes de redirection : PURL (Persistant URL) ex. : <http://purl.oclc.org/tuisp>
- Autres : ARK (Archival Resource Key) ex. : <http://catalogue.bnf.fr/ark:/12148/bpt6k85329c/f4.pagination>

Numérotation interne à l'application

- Respecte la définition et la structure d'une URI (Uniform Resource Identifier – défini par le W3C) : « préfixe : autorité nommante . nom »
- L'accès direct aux documents ou la possibilité de citation de ces documents depuis des systèmes d'information externes n'est pas une priorité.
- Possibilité de rejoindre ultérieurement le système DOI ou un équivalent.

Valeur caractéristique obtenue par hachage (mode de calcul) d'un fichier ou d'un lot de données.

Plusieurs algorithmes cryptographiques

- Message Digest 5 (MD5)
 - Conçu par Ronald Rivest en 1991 à partir du MD4
 - Empreinte numérique composée d'une séquence de 128 bits ou 32 caractères en notation hexadécimale : *5323e2cf888771d027ac3f2911183e99*
 - Probabilité très forte que, pour deux messages différents, leurs empreintes soient différentes.
 - Très populaire, mais n'est plus considéré comme un algorithme sûr – failles découvertes entre 1999 et 2004
- Secure Hash Algorithm 256 (SHA-256)
 - Conçu par la NSA (USA) en 2000 à partir du SHA-1
 - Calcul en 64 rondes avec 8 variables initialisées avec des constantes spécifiques concaténées à la fin des 64 tours (séquence de 256 bits) : *f5bc030e44e7c295780a495c0f2824a5adf205d6b137182b74b25feb1048d7ce*
 - est devenu le nouveau standard recommandé en matière de hachage cryptographique après les attaques sur MD5 et SHA-1

Les types de documents à archiver

Quel type de document est éligible à un archivage pérenne sur la plateforme PAC ?

- Tout document présentant une valeur patrimoniale scientifique ou technique
- De préférence des objets dits « primaires »
 - Documents originaux,
 - Bruts de scan, etc.
 - Si possible pas d'objets issus du retraitement d'objets primaires (ex. OCR, etc.)
- Issus d'archives définitives
 - La plateforme n'est pas un outil de diffusion

La modélisation OAIS des paquets d'information (rappel)

Trois types de paquet

- SIP – Submission Information Package
 - Paquet d'informations à verser
- AIP – Archival Information Package
 - Paquet d'information archivé
 - Produit à partir du SIP
- DIP – Dissemination Information Package
 - Paquet d'information diffusé
 - Produit à partir de l'AIP

La structure du document à archiver

Le document à archiver est composé de deux pièces

1. La description de l'archive

- Fichier sip.xml (schéma <http://www.cines.fr/pac/sip.xsd>)
- 3 sections décrivant :
 - Le document dans son projet d'archives
 - Le document proprement dit
 - Les fichiers du document



Document XML

2. Le dossier contenant les documents électroniques à archiver

- Répertoire « DEPOT »
- Sous-arborescence autorisée
- Tout fichier présent doit être décrit dans le fichier sip.xml

La structure du SIP, de l'AIP dans PAC

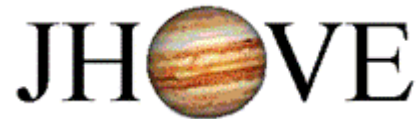
- Comprend
 - La ressource numérique
 - Les métadonnées de structure, administratives et descriptives
- Définie par un schéma XML
 - <http://www.cines.fr/pac/sip.xsd> pour le SIP
 - <http://www.cines.fr/pac/aip.xsd> pour l'AIP
- Le SIP est augmenté de métadonnées de préservation pour constituer l'AIP
 - date d'archivage du document,
 - Identifiant unique,
 - empreinte numérique (SHA-256) de chaque fichier
- Seul l'AIP est conservé

Les formats de documents à archiver

Formats identifiés et vérifiables :

- Format publié,
- Format largement utilisé,
- Format normalisé si possible

| Type | Format |
|-------|----------------------------|
| Texte | HTML, PDF, PDF/A, TXT, XML |
| Image | GIF, JPEG, TIFF, PNG |



Le système PAC est interfacé avec les outils Jhove et ImageMagick pour :



- Identifier,
- Valider,
- Caractériser,

Le format des fichiers transférés

L'identification, la validation et la caractérisation

Identification

Déterminer le format auquel se conforme un objet numérique.

Validation

Déterminer le niveau de conformité d'un objet numérique aux spécifications de son supposé format.

La validation se fait à deux niveaux, après quoi l'objet sera réputé bien-formé et valide

- Un objet est bien formé s'il suit les spécifications syntaxiques de son format
- Un objet est valide s'il est bien formé et suit des spécifications sémantiques supplémentaires

Caractérisation

Déterminer les propriétés spécifiques d'un objet numérique d'un format donné.

La caractérisation permet d'obtenir des informations de représentation (concept OAIS) telles que

- Le chemin ou l'URI
- La taille du fichier
- Le format, la version du format, le type MIME
- L'empreinte numérique (checksum SHA-1)

L'initiation d'un projet d'archives

Qui ? Tout organisme

- Produisant ou collectant en grande quantité des documents électroniques dont le contenu possède une valeur patrimoniale scientifique ou technique,
- Doté d'un système informatique pouvant être interfacé avec la plateforme PAC

Comment ? Deux phases

1. Phase préliminaire durant laquelle les points suivants sont abordés :
 - l'identification des informations à pérenniser
 - la liste des données et métadonnées transmises au CINES (format, taille, nombre...)
 - l'analyse de faisabilité (sécurité, aspects légaux, coûts et risques...) ;
 - l'évaluation de la volumétrie et des ressources requises.
2. Phase dite de définition
 - la définition précise des objets à transférer
 - les termes et conditions du protocole de transfert (restrictions d'accès, communicabilité au public)
 - la planification des transferts physiques ;
 - la formation du personnel du service versant à l'utilisation du système PAC

Le producteur

- Personne physique ou morale, publique ou privée, qui a produit, reçu et conservé des archives dans l'exercice de son activité.

Le service versant

- Organisation qui transfère une archive à un service d'archives

Le service de contrôle

- Personne physique ou morale qui effectue le contrôle scientifique, juridique et technique des documents archivés, et éventuellement valide les demandes de communication d'archives

Le service d'archives

- Organisation recevant le document à archiver transféré et chargée de la conserver pour permettre à une communauté d'utilisateurs/un service demandeur d'y accéder et de l'utiliser

L'utilisateur

- Toute personne ou système client en relation avec le service d'archives pour trouver les informations archivées présentant un intérêt, et pour accéder au détail de ces informations, dans le respect de la législation applicable en matière de communication des archives.

Transfert d'archives

- Transmission physique d'une archive ou d'un ensemble d'archives par un service versant à un service d'archives

Modification d'archives

- Modification des métadonnées et/ou du document pour en assurer la préservation

Elimination d'archives

- Elimination des métadonnées et/ou du document à la demande du services d'archives, du service versant ou du service de contrôle

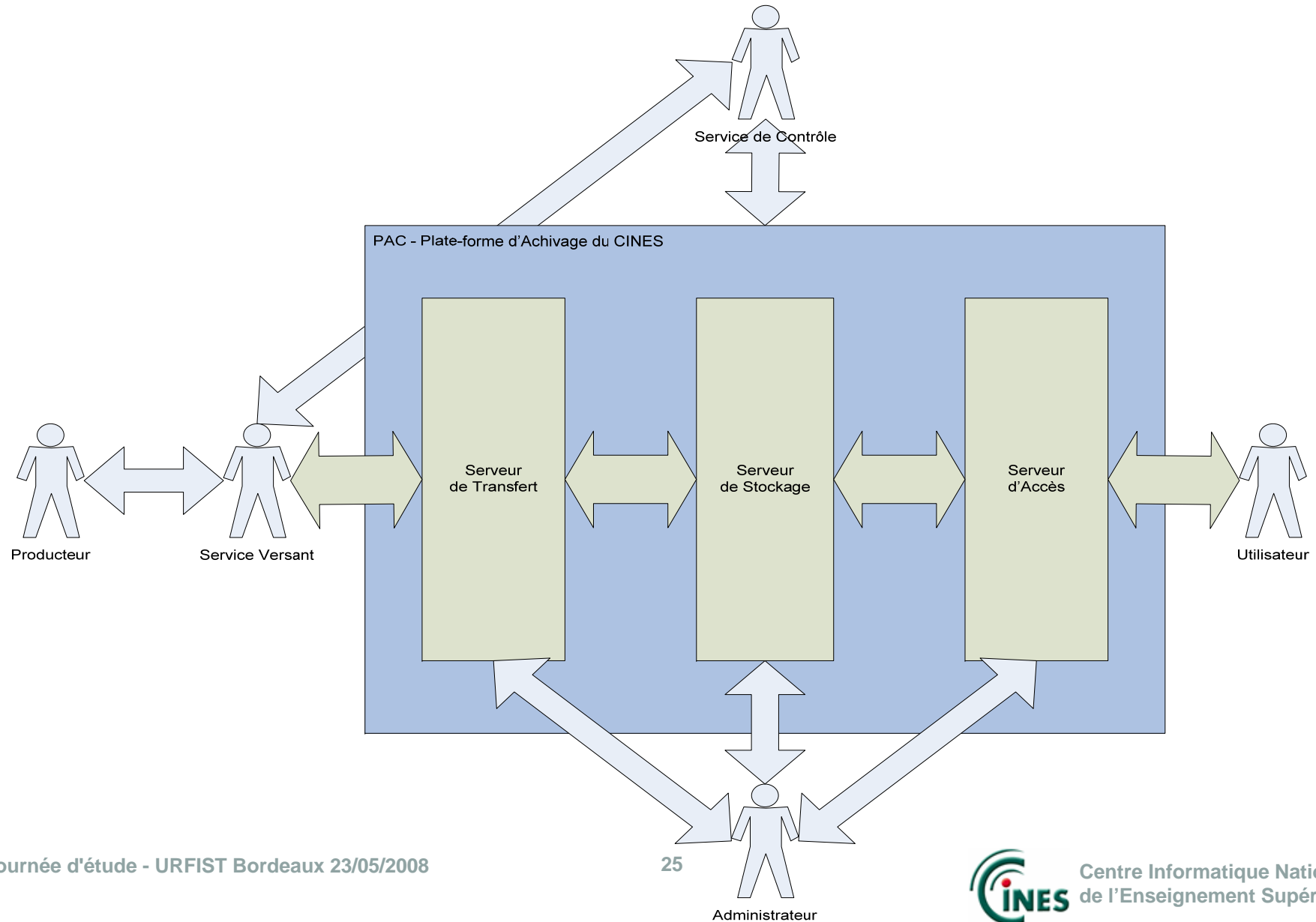
Restitution d'archives

- Transmission de documents par le service d'archives au service versant ou au producteur afin de leur en restituer la garde

Communication d'archives

- Transmission de copie de document à un utilisateur ayant l'autorisation du service versant et /ou du service de contrôle

L'architecture logique de la plateforme



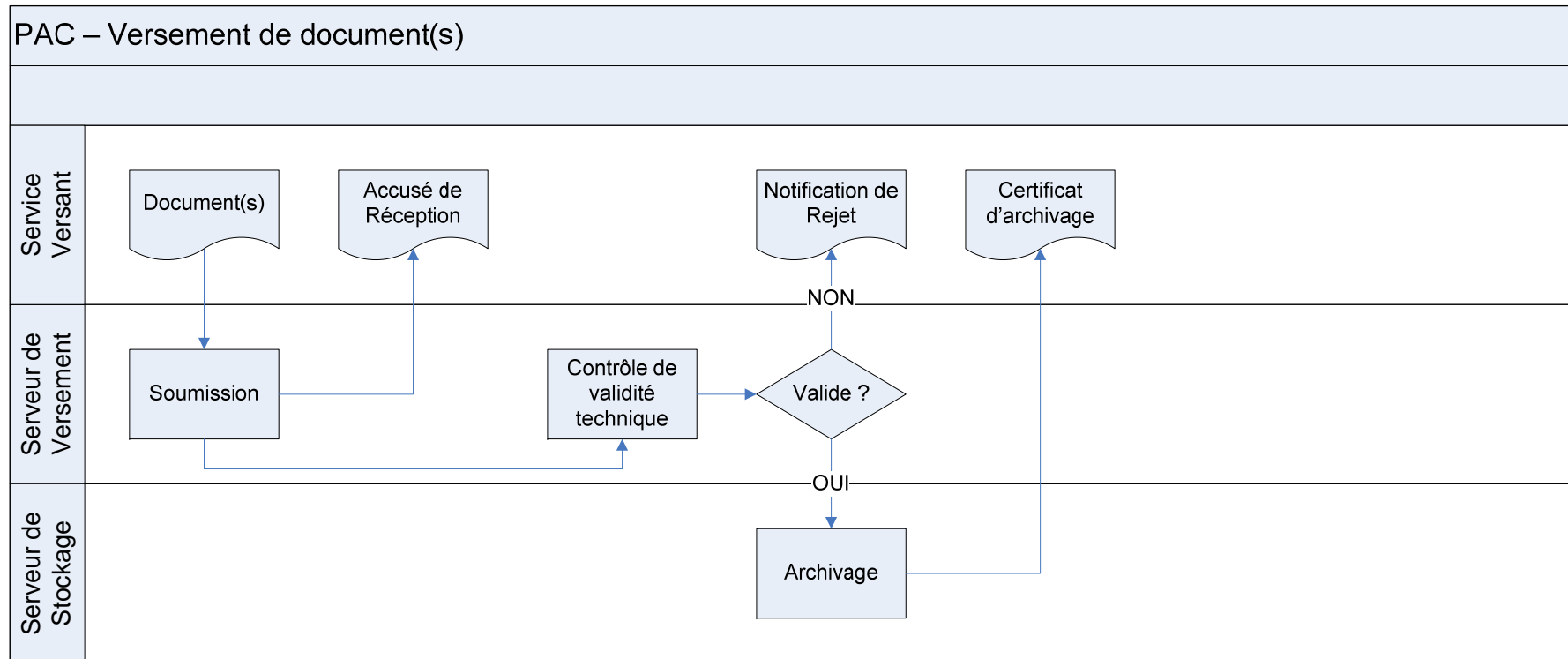
Les principes de fonctionnement

| | | |
|------------------|--------------------------|--|
| Transfert | réception des SIP | <i>détection d'un nouveau transfert envoi d'un accusé de réception</i> |
| | contrôle des SIP | <i>structure informatique conformité des métadonnées sip.xml par rapport au schéma sip.xsd correspondance entre la description sip.xml et les fichiers qui composent le document contrôle et validation du format des fichiers calcul de l'empreinte numérique de chaque fichier</i> |
| | création des AIP | <i>création de l'identifiant du document archivé mise à jour des métadonnées : sip.xml > aip.xml transfert de l'AIP au serveur de stockage</i> |

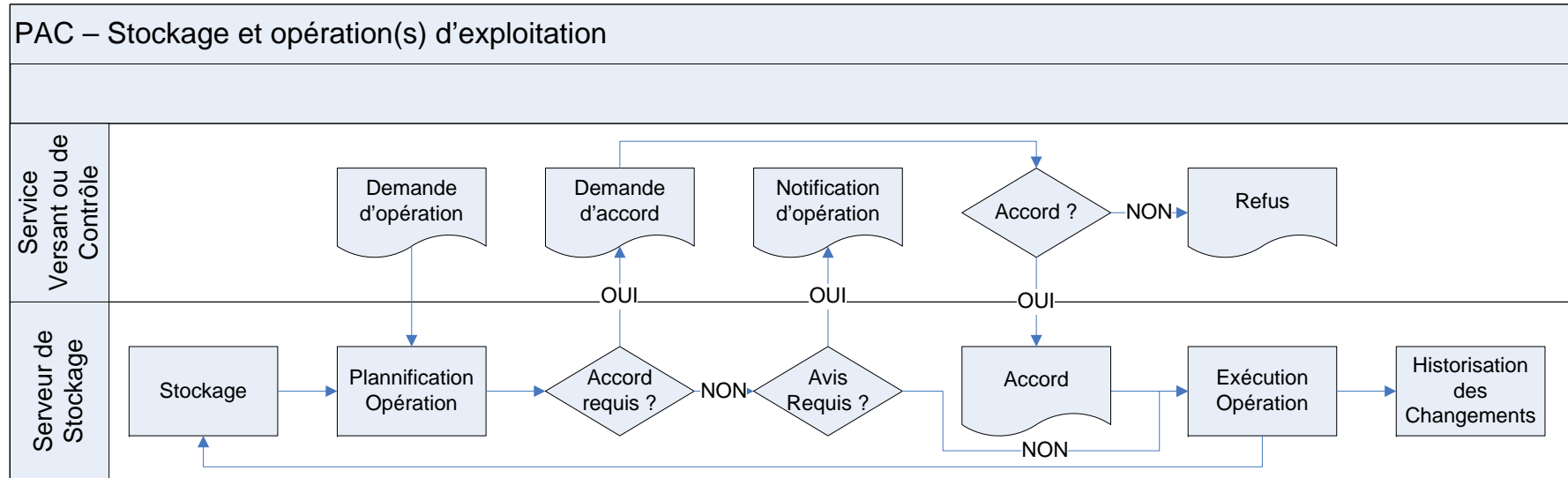
| | | |
|-----------------|--|--|
| Stockage | archivage des AIP | <i>copie multiple de l'AIP sur les différents médias ou supports envoi du certificat d'archivage</i> |
| | vérification périodique de l'intégrité des AIP archivés | |
| | migration technologique | |
| | fourniture d'états et de statistiques | |

| | |
|--------------|--|
| Accès | contrôle de l'authentification de l'utilisateur |
| | consultation du catalogue des AIP archivés |
| | communication d'une copie d'un document archivé |

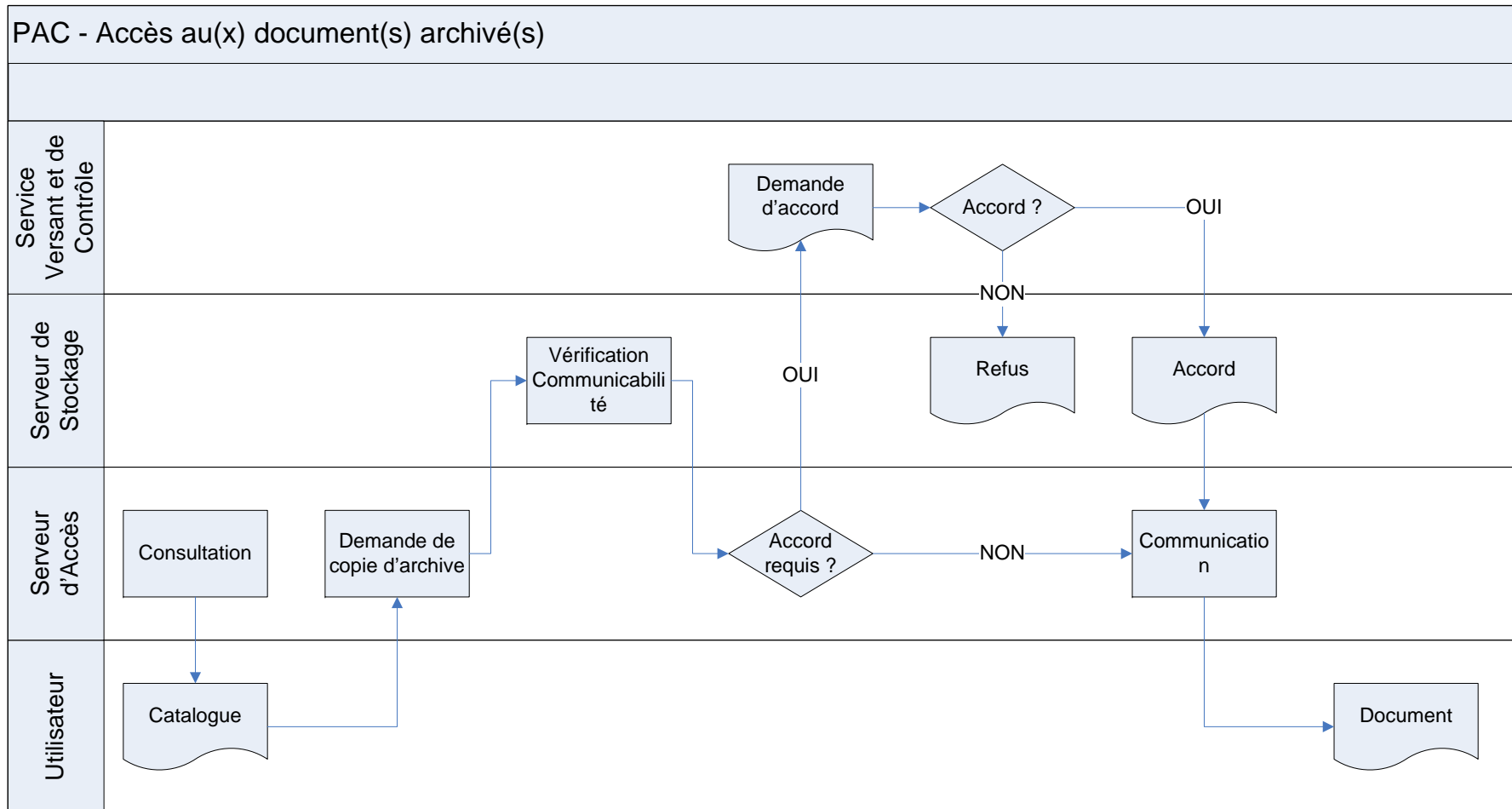
Les étapes du versement d'archives



Les étapes du stockage d'archives



Les étapes de la communication d'archives



- Phase 1 : Développement interne d'une première plateforme pour valider les services attendus sur le projet ABES/STAR, capacité de stockage réduite (300 Go) – PAC v1.0
 - Basée sur les standards du domaine
 - Modèle OAIS
 - Protocole standard d'échange de données pour l'archivage
 - Métadonnées Dublin Core
 - Liste des formats de fichier acceptés volontairement limitée
 - Formats publiés, largement utilisés, normalisés si possible
 - HTML, PDF, TXT, XML
 - GIF, JPEG, TIFF, PNG
 - Architecture basée sur les logiciels libres
 - Java, PostgreSQL, Jhove, ImageMagick
 - Premiers tests d'archivage des thèses en Mars 2007
 - Début de l'exploitation en production fin 2007

- Phase 2 : Appel d'offres notifié fin 2007 pour l'acquisition d'une plateforme de stockage capable de gérer de larges volumes (20 To extensibles à 40To) – PAC v2.0
 - Basée sur les standards du domaine
 - Modèle OAIS
 - Protocole standard d'échange de données pour l'archivage
 - Métadonnées Dublin Core
 - Liste des formats de fichier acceptés identique à PAC v1.0
 - Architecture basée sur du matériel SUN, le logiciel Arcsys et des logiciels libres
 - Java, MySQL, Jhove, ImageMagick
 - Premiers tests de versement et de migration (documents archivés sur PAC v1.0) en Mars 2008
 - Début de l'exploitation fin du deuxième trimestre 2008

Le projet d'archives « Thèses électroniques » ABES

Initié suite à l'arrêté du 7 août 2006 relatif aux modalités de dépôt, de signalement, de reproduction, de diffusion et de conservation des thèses ou des travaux présentés en soutenance en vue d'un doctorat

- Les doctorants déposent leur thèses au format électronique dans la BU (Bibliothèque Universitaire) de leur lieu de soutenance
- Les BU versent les thèses électroniques à l'ABES via l'outil STAR
- Après 3 étapes de validation, les thèses éligibles à l'archivage sont transférées sur la plateforme PAC
- Une copie est éventuellement disponible en ligne pour la communauté des internautes sur un site de diffusion
- L'outil STAR centralise les demandes de communication d'archives

Ce projet est actuellement en phase de production, l'archivage des thèses a débuté fin 2007

<http://www.abes.fr/abes/page,555,manuel-de-lutilisateur-star.html>

Le projet d'archives « revues SHS » Persée

Initié en 2006 pour répondre à un projet de numérisation massive et de préservation de collections rétrospectives de revues en Sciences Humaines et Sociales par l'équipe Persée (Université Lumière – Lyon 2).

- La chaîne de numérisation assure
 - une digitalisation de masse
 - une centralisation et une robotisation des traitements
 - un archivage pérenne des données
- La chaîne de documentation comprend
 - un outil de description des collections
 - des outils de documentation et de suivi
 - des étapes de contrôle qualité et de validation
 - des outils pour la diffusion des données générées (<http://www.persee.fr/>)

Ce projet est actuellement dans sa phase préliminaire (définition des objets à archiver, du protocole de versement). Les premiers tests d'intégration sont planifiés pour le deuxième trimestre 2008

Les projets en cours d'étude

Les projets suivants sont en cours d'étude

1. Archivage pérenne des documents versés dans les Archives Ouvertes (HAL – Hyper Article en Ligne, <http://www.archives-ouvertes.fr/>)
 - Étude préliminaire en cours
 - Calendrier prévisionnel fin 2008
2. Autres projets de numérisation et d'archivage proposés
 - IRHT
 - TGE-Adonis
 - CERIMES
 - EFEO

Tous les projets d'archives partageront la même plateforme

- Mutualisation de l'infrastructure matérielle d'archivage
- Protocole de versement générique
- Diminution des coûts de mise en place et d'exploitation

La mise en place d'une stratégie pour la préservation à long terme de documents électroniques issus de l'IST en est à ses débuts.

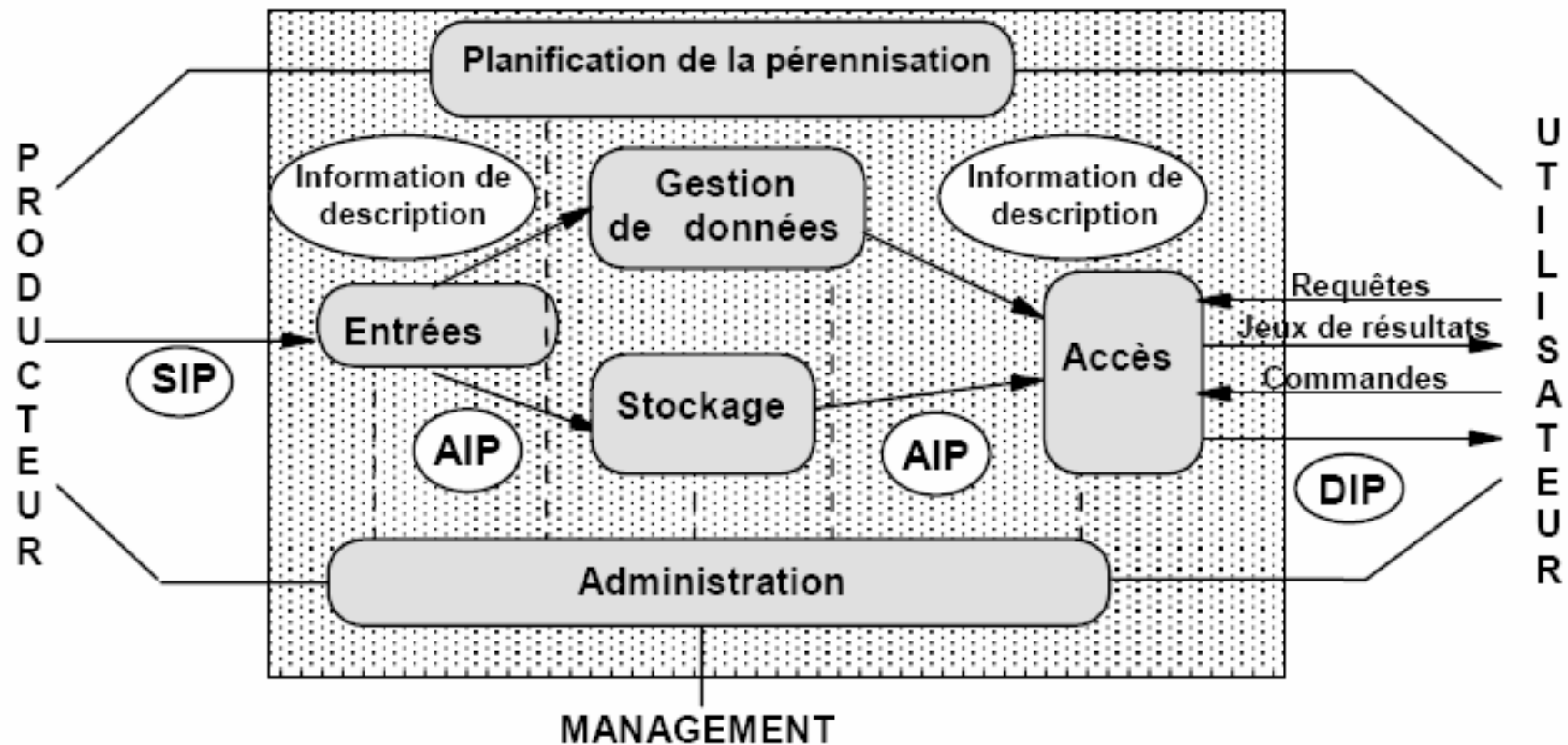
Elle réussira si la chaîne fonctionnelle d'archivage est claire et les acteurs identifiés. Le CINES y jouera un rôle clé dans la mesure où :

- Il est maintenant un acteur reconnu du domaine de la préservation à long terme de documents numériques :
 - Expérience dans l'archivage pérenne (documents nativement produit au format électronique ou issus de la numérisation de publications, photos, etc.)
 - Une étude récente menée par le TGE-Adonis préconise une collaboration étroite IN2P3 – CINES pour l'accès et la préservation de l'information SHS
 - Rôle confirmé par le ministère de tutelle
- Il reçoit de nombreuses sollicitations de laboratoires et d'universités pour divers services :
 - Aide et conseil dans la construction de projets d'archivage à long terme, retours d'expérience, propositions de projets d'archives
 - Dans le respect du contexte législatif archivistique français



Annexes

Le modèle fonctionnel O AIS (rappel)



Pour en savoir plus : le groupe de travail PIN

- PIN (pérennisation de l'information numérique) groupe de travail de l'association Aristote
- Lieu de rencontre et d'échanges entre informaticiens, archivistes et bibliothécaires
- Principalement animé par le CNES (Claude Huc), la BnF et la DAF
- Réunions trimestrielles (environ 30 participants réguliers)
- Un site web : <http://pin.cnes.fr>
- Une formation spécialisée (2 sessions par an)