L'archivage pérenne du document numérique au CINES

CINES (O.Rouchon)

Rencontres RNBM – 3 Octobre 2007

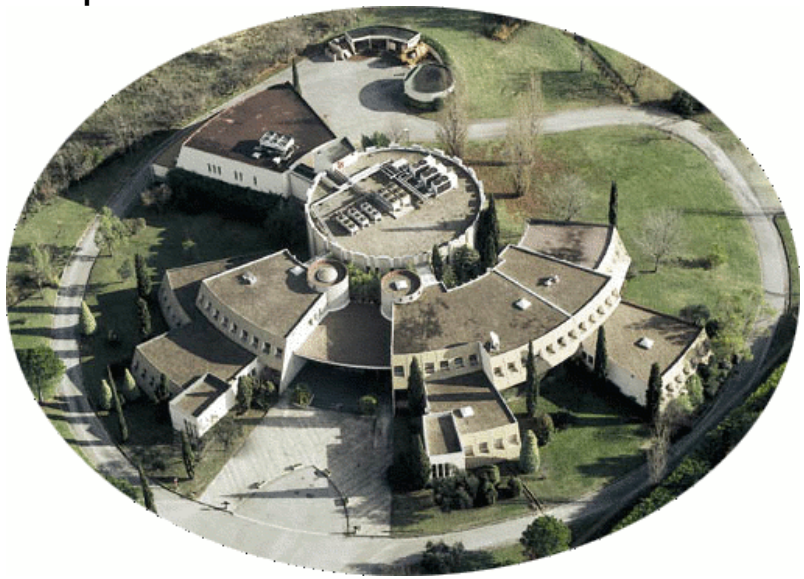
Sommaire

- La mission d'archivage du CINES
- Le contexte, la problématique et les constats
- Les défis, orientations et choix pour l'archivage au CINES
- Les types de documents à archiver
  - Le cycle de vie des documents
  - L'initiation du projet d'archives
  - La structure du document à archiver
- Les acteurs
- L'architecture logique de la plateforme
- Les échanges
  - Les principes de fonctionnement
  - Les étapes des différents échanges
- L'état des lieux

Le CINES

Centre Informatique National de l'Enseignement Supérieur

- Basé in Montpellier (Hérault, France)
- Créé en 1999, succédant au CNUSC (Centre National Universitaire Sud de Calcul) – créé in 1980
- Placé sous la tutelle de la DGRI (Direction Générale de la Recherche et de l'Innovation) du Ministère de l'Enseignement Supérieur
- Principales missions
  - Calcul numérique intensif,
  - Exploitation de base de données et d'applications
  - Expertise et formation en matière de réseaux informatique (avec RENATER)
- Plus d'information : http://www.cines.fr/



La mission d'archivage du CINES

Depuis 2004, le CINES a une mission nationale d'archivage du patrimoine scientifique.

- 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

- Pour la remplir, le CINES a mis en place le projet PAC, qui vise à doter le CINES d'une plate-forme et d'un service d'archivage numérique pérenne
- L'équipe : 1 chef de projet, 4 ingénieurs, 1 archiviste (5 ETP)
- 2 projets pilotes
  - Archivage des thèses électroniques
  - Archivage des revues SHS du portail Persée
- Contraintes
  - Solution générique, basée sur les standards pour une évolution,
  - Facilité de veille technologique et de migration

Le contexte, la problématique et les constats

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 :

- Connaissance perdue du contenu des fichiers,
- Format de fichier inconnu,
- Support physique détérioré,
- Logiciel ou matériel de lecture disparu

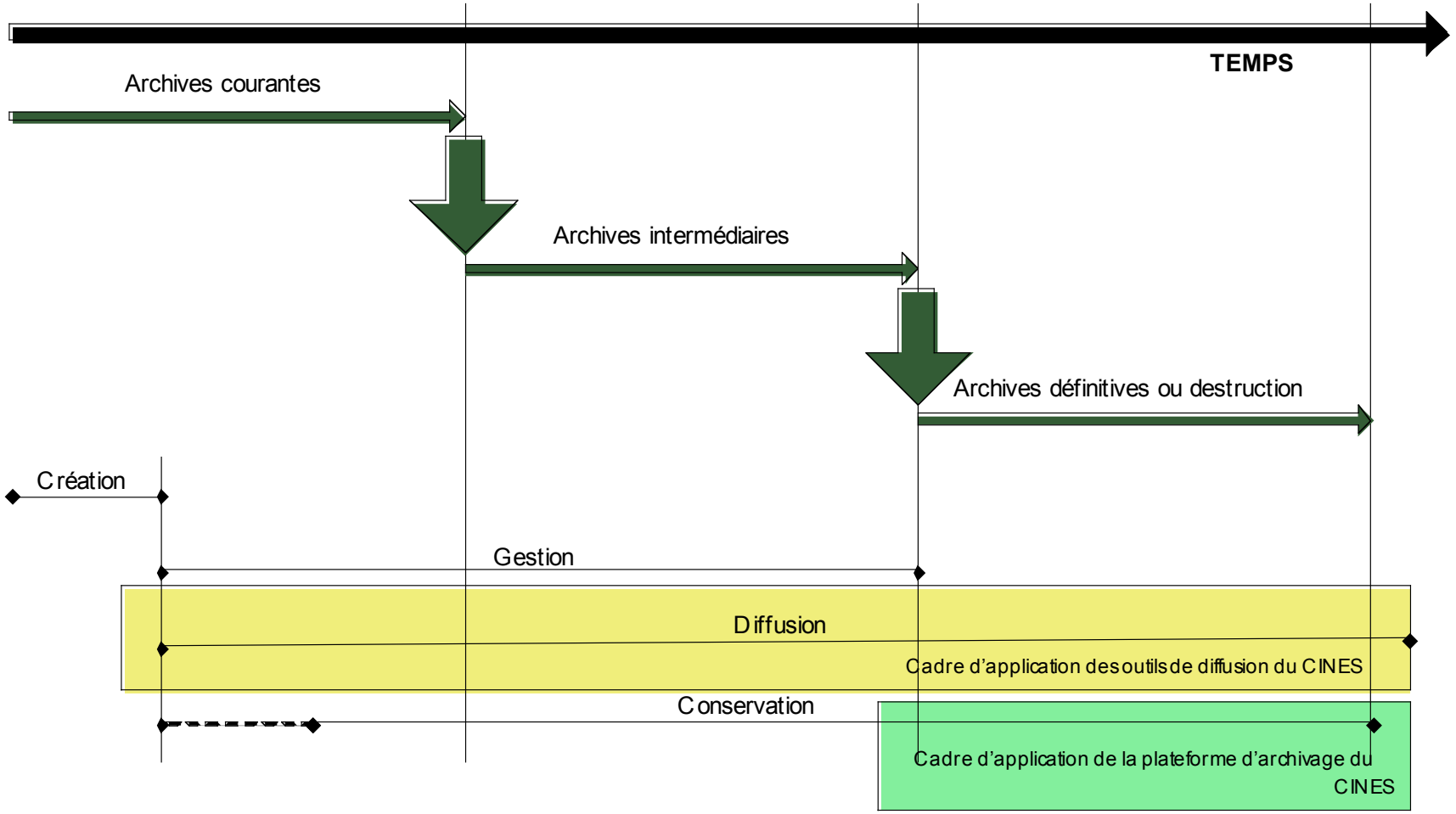
Les défis, orientations et choix pour l'archivage au CINES

Contrainte	Solutions
Connaissance du contenu	• Utilisation de métadonnées • Identification unique et pérenne des documents archivés
Format de fichier inconnu	• Privilégier les formats durables • Identification, validation des formats • Migration logique
Support physique détérioré	• Gestion du vieillissement des médias • Migration physique
Logiciel ou matériel de lecture disparu	• 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.
- Standard d'échanges de données pour l'archivage électronique, versement, communication, élimination
  - DAF, DGME, version 1.0, mars 2006.
- Normes internationales de description archivistique
  - ISAD-G – international standard for archival description, general
  - ISAAR-CPF – international standard archival authority record, corporate bodies, persons & families
- Métadonnées descriptives de l'archive
  - DCMI – Dublin Core Metadata Initiative
  - METS – Metadata Encoding and Transmission Standard
- Identifiant unique et pérenne
  - Interne, séquentiel, basé sur le principe URI
  - Pas d'utilisation pour le moment d'identifiant persistant externe de type DOI, ARK
- Empreintes numériques
  - Hashing MD5, SHA-256

Le cycle de vie des documents



Les types de documents à archiver

- Présentant une valeur patrimoniale scientifique ou technique
- De préférence des objets dits « primaires »
  - Documents originaux,
  - Bruts de scan, etc.
- Issus d'archives définitives
- Dans un format identifié et vérifiable :
  - Format publié
  - Format largement utilisé (ou promis à l'être)
  - Format normalisé si possible

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

Le système PAC est interfacé avec l'outil Jhove pour

- Identifier, Valider, Caractériser,

Le format des fichiers transférés

L'identification, la validation et la caractérisation

Fonction	Description
Identification	**Déterminer le format auquel se conforme un objet numérique.** *Répondre à la question: « J'ai un objet numérique, dans quel format est-il ? »*
Validation	**Déterminer le niveau de conformité d'un objet numérique aux spécifications de son supposé format.** *Répondre à la question: « J'ai un objet numérique qui est censé être un PDF. En est-il un ? »* **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** **Exemple – un fichier TIFF est bien-formé si son entête est composée de 8 octets suivis d'une séquence IFD (Image File Directories), et valide si (dans le cas d'un fichier RGB) il a 3 valeurs par pixel**
Caractérisation	**Déterminer les propriétés spécifiques d'un objet numérique d'un format donné.** *Répondre à la question: « J'ai un objet numérique au format PDF, quelles sont ses propriétés ? »* **La caractérisation permet d'obtenir des information 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)**

La structure du document à archiver

Document à archiver 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
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

Les acteurs

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

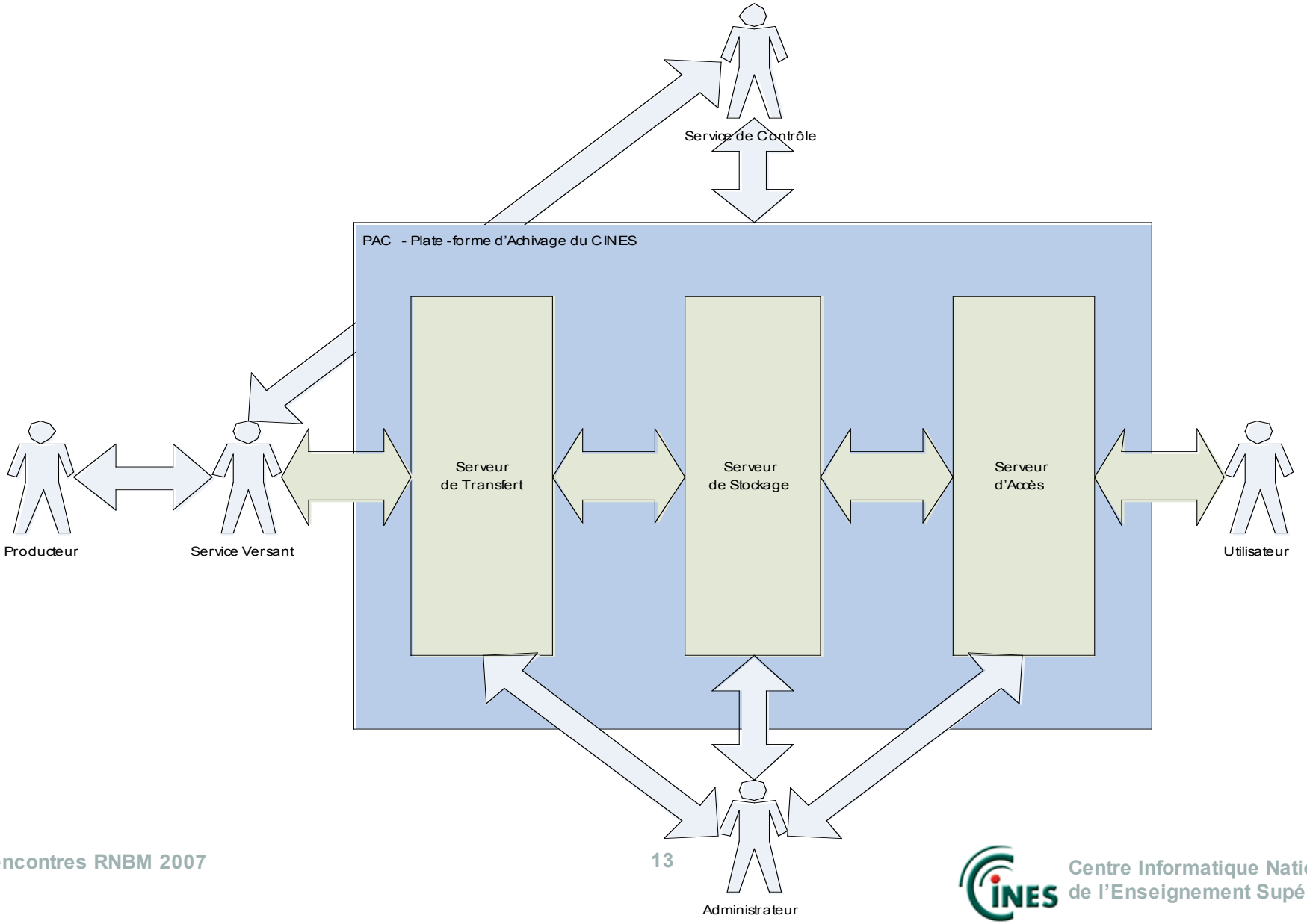
Le service d'archives

- 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.

L’architecture logique de la plateforme



Les échanges

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

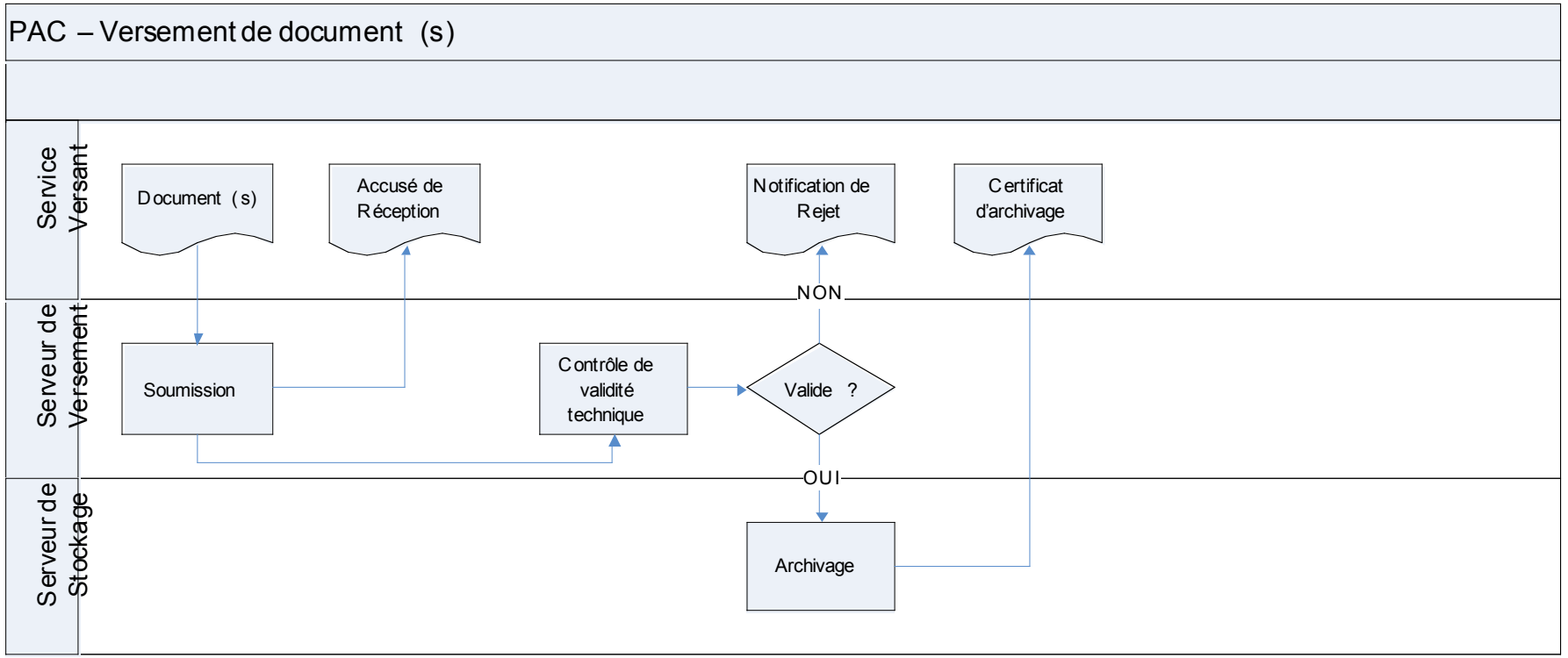
- Transmission de copie de document à un utilisateur ayant l'autorisation du service versant et /ou du service de contrôle

Les principes de fonctionnement

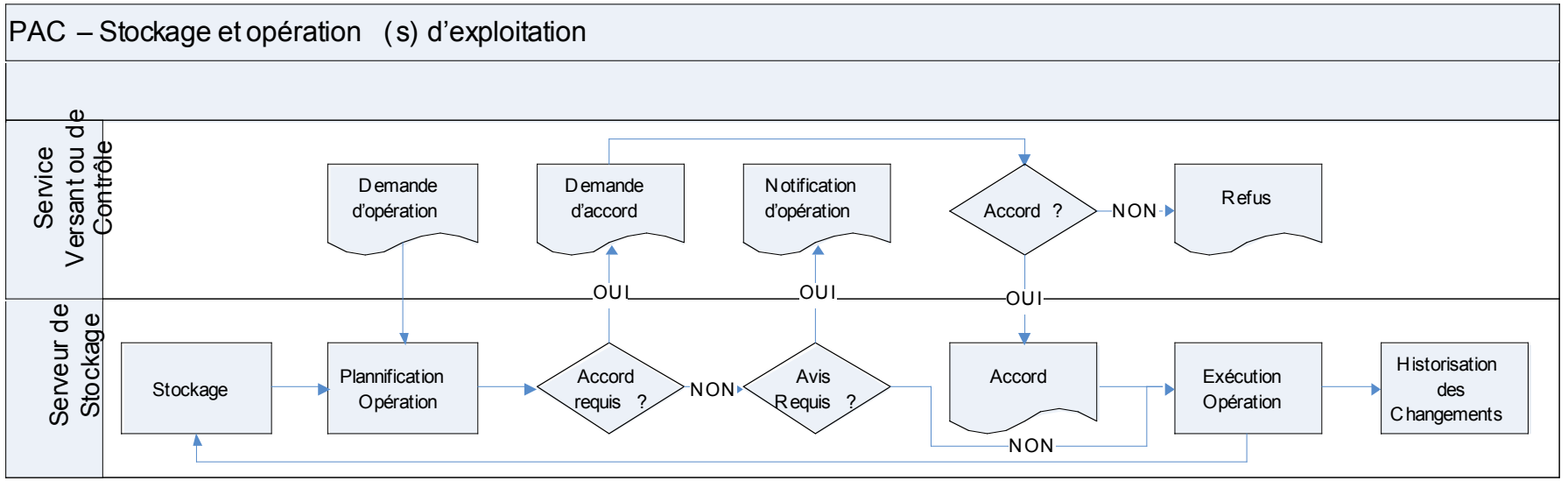
Trois serveurs logiques : transfert, stockage, accès

Serveur	Fonctions	
Transfert	**réception des SIP**	*détection d'un nouveau transfert* *envoi d'un accusé de réception*
	contrôle des SIP	*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*
	création des AIP	*calcul de l'empreinte numérique de chaque fichier* *création de l'identifiant du document archivé* *mise à jour des métadonnées : sip.xml > aip.xml* *transfert de l'AIP au serveur de stockage*
Stockage	**archivage des AIP**	*copie multiple de l'AIP sur les différents médias ou supports* *envoi du certificat d'archivage*
	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	

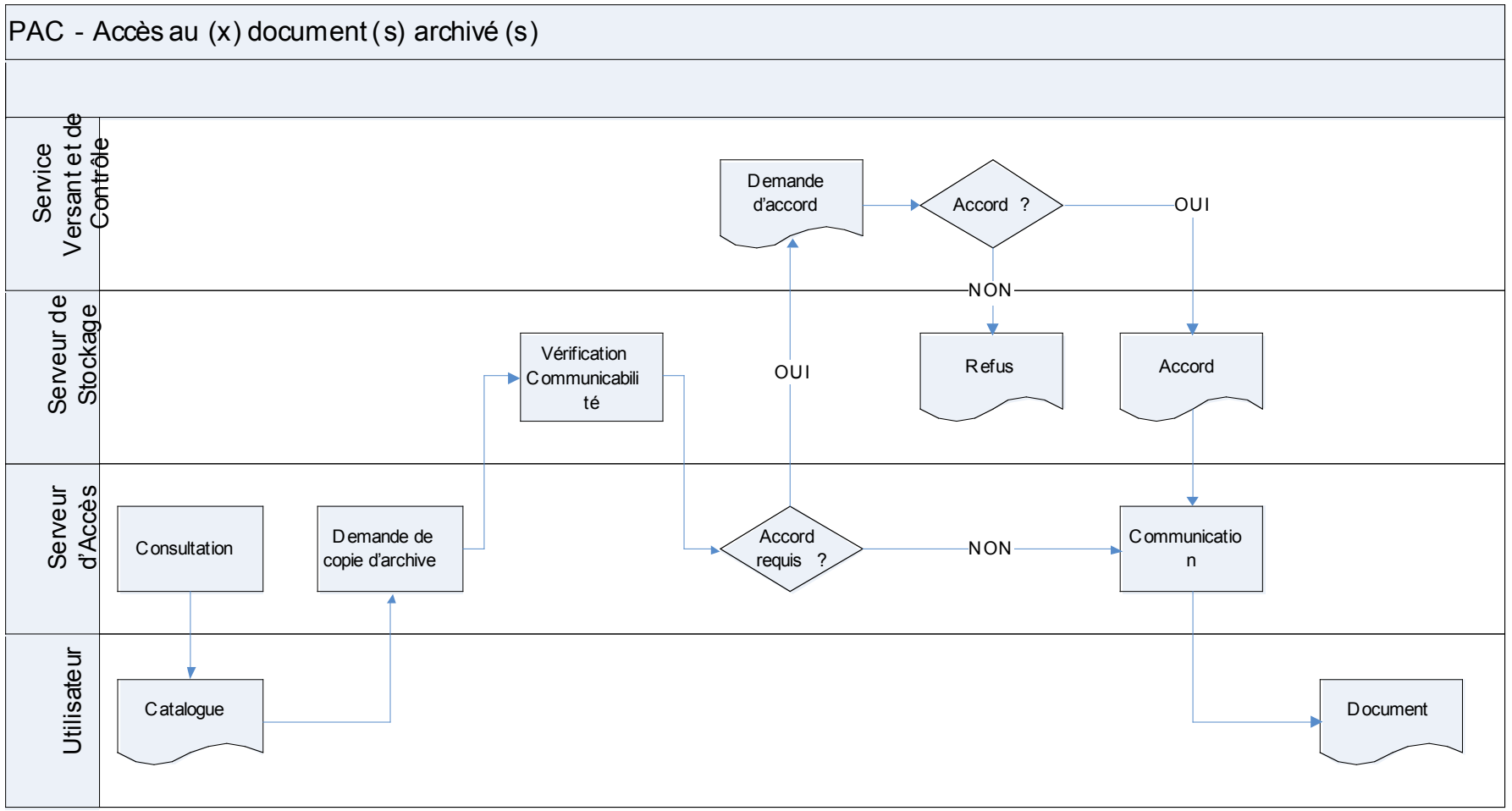
Les étapes du versement d'archives



Les étapes du stockage d'archives



Les étapes de la communication d'archives



Etat des lieux (Juin 2007)

- Phase 1 : Développement interne d'une première plate-forme pour valider les services attendus sur les 2 projets pilotes – premiers tests d'archivage des thèses en Mars 2007, déploiement en production fin Juin 2007

Protocole transfert	Soumission de SIP sur serveur de transfert via SFTP
Langage	Composants des serveurs développés en Java
Base de données	Base de données PostgreSQL
	Utilisation de bibliothèques OpenSource (module d'analyse de fichiers, calcul des empreintes numériques, etc.)
	Déclenchement des routines (accusé de réception, vérification, archivage) à intervalles réguliers – traitement par lots
Interface utilisateur	administration en mode web

- Phase 2 : Appel d'offres en cours pour l'acquisition fin 2007d'une plate-forme de stockage capable de gérer de larges volumes (40 To)

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)

Annexes

Termes et définitions

Terme	Définition
Archives courantes	Les archives courantes regroupent les documents nécessaires à l'activité des services qui les ont produits et conservés pour le traitement de leurs affaires courantes.
Archives intermédiaires	Les archives intermédiaires ne sont plus utilisées mais restent utiles et doivent être conservées temporairement (besoins administratifs ou juridiques). À l'issue de cette durée de conservation, les archives intermédiaires font l'objet d'un tri et sont soit conservées définitivement soit éliminées.
Archives définitives	Les archives définitives ont vocation à être conservées pour des raisons historiques, juridiques ou patrimoniales.
Diffusion	Dans le langage courant, le terme diffusion fait référence à une notion de « distribution », de « mise à disposition » (diffusion d'un produit, d'une information). La diffusion de documents électroniques consiste à faciliter leur disponibilité et leur accès à court ou moyen terme.
Conservation	D'une manière générale, la conservation est l'acte qui consiste à préserver un élément dans un état constant.
Archivage pérenne	L'archivage pérenne consiste à conserver le document électronique, le rendre accessible, en préserver l'intelligibilité, le tout sur le très long terme, trente ans et au-delà.