Pil@e

Projet de mise en œuvre des concepts-piliers de l'archivage numérique

Plate-forme pilote d'archivage électronique

Réunion du groupe PIN

23 avril 2007

Françoise Banat-Berger (DAF) et Gabriel Ramanantsoavina (DGME)

Le projet PILAE : origines

- En 2005, la DAF fait réaliser une étude sur les coûts d 'une plate-forme d 'archivage électronique (société Parker-Williborg)
- Le scénario choisi était le suivant :
  - sur la base d 'outils existants, paramétrage d 'une solution de plate-forme d 'archivage électronique à faire réaliser en tant que pilote au niveau national
  - pour ré-utilisation du modèle, voire des outils, par les services qui désireraient développer leur propre plate-forme (au sein des services producteurs, au sein des services d 'archives)

Le projet PILAE : origines

- En 2005 et 2006, la DAF et la DGME développent plusieurs référentiels :
  - avec la DCSSI, l 'élaboration d 'une politique d 'archivage (conditions d 'ordre juridique, technique, archivistique, organisationnelle, fonctionnelle) nécessaires pour que des documents et données conservent leur force probante d 'origine durant le processus d 'archivage
  - avec la DGME, élaboration du standard d 'échange de données pour l 'archivage (soit un format de métadonnées orienté échange entre services producteurs, versants et services d 'archives)

Le projet PILAE : origines

- L 'enjeu du projet était notamment l 'utilisation du pilote au sein des Archives nationales (site de Fontainebleau) dans le cadre du service Constance durant la période transitoire 2008-2011 avant l 'ouverture du nouveau centre des Archives nationales de Pierrefitte sur Seine et de sa plate-forme d 'archivage électronique
- Il s 'agit dans le cadre du développement de l 'administration électronique de pouvoir accueillir, traiter, conserver et communiquer les archives nativement numériques produites par les services centraux de l'Etat et par conséquent de permettre une certaine automatisation des tâches et une meilleure sécurité de la conservation

Le projet PILAE : origines

- Parallèlement, au sein du conseil général des Yvelines, développement en interne d 'outils matériels et logiciels visant :
  - à recevoir, contrôler et prendre en charge les archives du contrôle de légalité dématérialisé (à partir de la plate-forme de télétransmission FAST qui développe de son côté un module d 'export au format du standard d 'échange de données pour l 'archivage). La plate-forme devra pouvoir dans un second temps recevoir d 'autres types d 'archives
  - à rechercher et communiquer ces archives
  - l 'infrastructure de stockage dans un premier temps est celle déjà utilisée pour stocker les fonds d 'archives papier qui ont été numérisés pour mise en ligne sur internet
  - les outils ainsi développés seront mis à disposition en open-source

Le projet PILAE : origines

- Les enjeux étaient notamment :
  - de tester le standard d 'échange de données pour l 'archivage (ou protocole standard d 'échanges PSE)
  - de tester le traitement de plusieurs natures d 'archives numériques (données extraites de bases de données, documents issus de GED et décrits par une base de donnée, messageries électroniques, flux de données sécurisées…)
  - de tester la mise en œuvre de contrats entre services versants, producteurs et d 'archives, conditionnant un certain nombre de vérifications automatiques
  - de tester des conversions de formats en entrée du système par lots
  - de tester les mécanismes de contrôles d 'intégrité, d 'horodatage
  - de tester les mécanismes de réplication

Le projet PILAE : origines

- Les points délicats
  - organisationnels :
    - une équipe projet restreinte (DAF/DGME)
    - un DSI au départ peu motivé, ne connaissant pas le métier, peu au fait des nouvelles technologies
    - une équipe Constance très qualifiée mais connaissant surtout la problématique de l 'archivage des bases de données
    - des archivistes dans les ministères généralement peu au fait de l 'archivage électronique

Le projet PILAE : origines

  - techniques
    - implémentation encore inexistante du standard d 'échange dans les applications métier en amont : l 'application PILAE a dû intégrer une partie cliente pour la préparation des versements (on joue le rôle des services versants et producteurs) et concrètement développer à la place des producteurs des moulinettes pour formater les versements
    - l 'aspect novateur de plusieurs fonctionnalités et technologies (contrôle d 'intégrité, signature électronique, standard d 'échange, conversion de formats, réplication)

Le projet PILAE : le marché

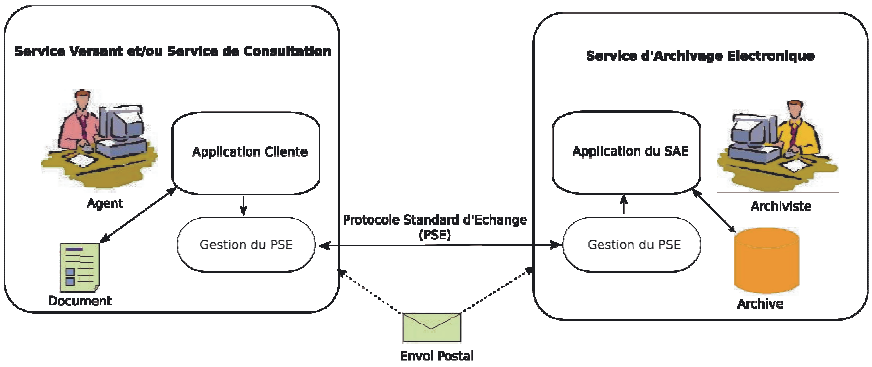
- Il a été attribué en septembre 2006 à la société Cecurity.Com éditrice du coffre-fort électronique communicant (CFEC) avec en sous-traitance IBM pour la partie infrastructure de stockage
- Le savoir-faire documentaire n 'était pas le cœur du projet :
  - tout ce périmètre sera largement pris en charge par le futur système d 'information des Archives nationales qui devrait à terme opérer pour la partie recherche et consultation tant pour les archives papier que pour les archives électroniques

Le projet PILAE : le marché

- Deux tranches :
  - une ferme qui s 'est achevée au début du mois de mars 2007 (réalisation des spécifications fonctionnelles détaillées)
  - une conditionnelle (aujourd 'hui affermie) correspondant à la réalisation proprement dit
  - un calendrier qui s 'achève en avril 2008 (correspond à la validation du service régulier VSR)
- Une organisation classique :
  - COPIL et équipe de projet
  - un comité utilisateurs pour la tranche de la réalisation (les services qui fournissent des données à tester dans le cadre du projet, les archivistes en poste dans les ministères) pour réagir sur la maquette, la cinématique des écrans, pour participer aux tests utilisateurs
  - une conduite de projet informatique classique, méthode itérative (UML)

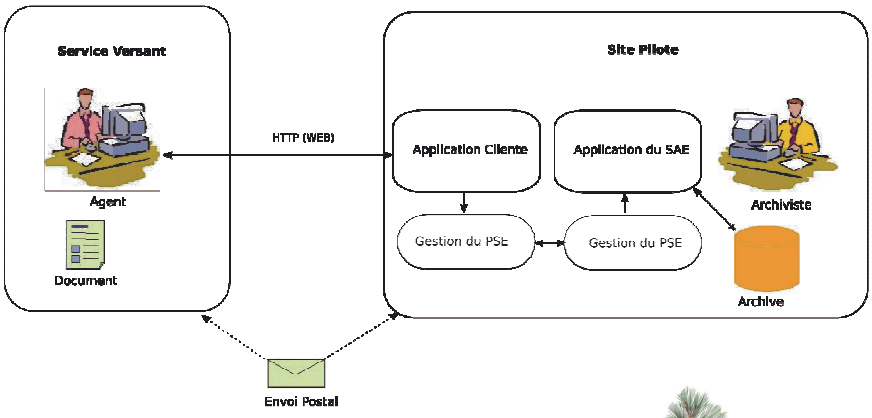
Architecture applicative : le principe du standard d'échange

Système cible

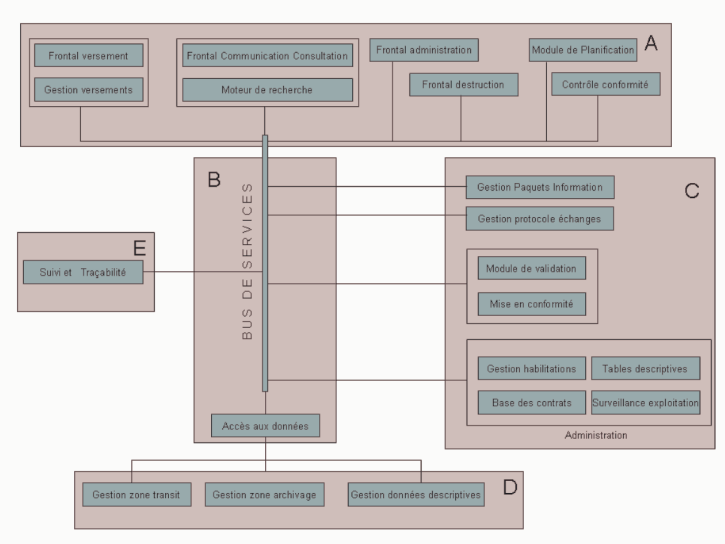


Architecture applicative : le principe du standard d'échange

Système pilote



Architecture applicative : le principe SOA



Modélisation des paquets d'information

- Norme OAIS : 3 types de paquet
  - SIP (*Submission Information Package*) : paquet d'informations à verser
  - AIP (*Archival Information Package*) : paquet d'informations archivé
  - DIP (*Dissemination Information Package*) : paquet d'information diffusé
- Modélisation à appliquer au pilote (et au PSE)

Modélisation des paquets SIP

Paquet SIP

- Document XML + fichiers joints ~ (PSE : ArchiveTransfer)
- Conversion des documents
- Complété et modifié par archiviste
- Conservé en zone de transit
- Pas de recherche documentaire sur les paquets SIP
- Un identifiant unique par paquet SIP

Modélisation des paquets AIP

Paquet AIP

- Production après validation archiviste à partir d'un SIP reçu (conservé pour traçabilité)
- Paquet AIP ~ PSE : Objet Archive complet
- Lien AIP – SIP maintenu
- Liens entre AIP possibles
- Un identifiant unique par paquet AIP
- L'AIP contient les 2 dernières versions des documents convertis
- Recherche via moteur + données descriptives

Modélisation des paquets DIP

Paquet DIP

- Production suite recherche pour consultation
- Sélection du contenu par le chercheur
- Conservé en zone de transit
- Un identifiant unique par paquet DIP
- Le chercheur consulte au format le plus récent
  - Sauf demande particulière de sa part

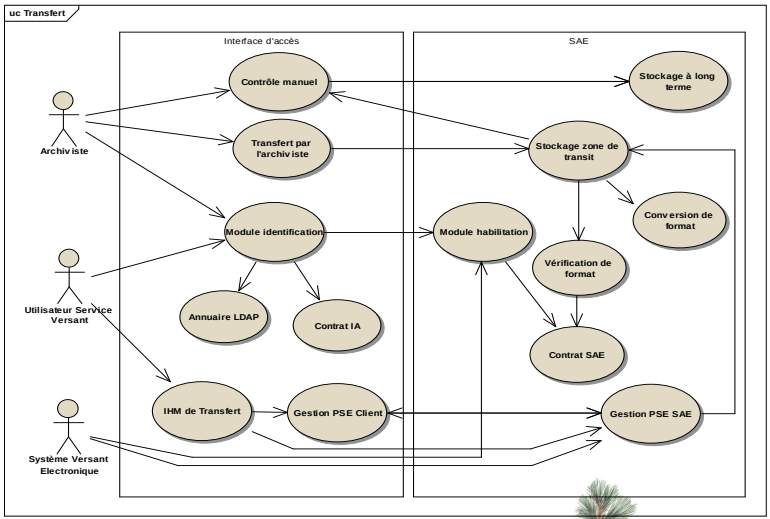
Le volet fonctionnel du projet PIL@E

- Présentation synthétiques d'écrans représentatif
- Modélisation des paquets d'information : API, SPI & DPI

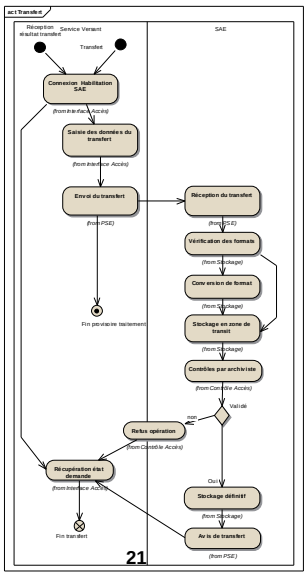
La démarche

- Les cas d'utilisation « fonctionnel »
  - Transfert
  - Communication
  - Modification
  - Destruction
- Les cas d'utilisation « acteur »
  - Archiviste
  - Utilisateur d'un service versant ou producteur
  - Grand public

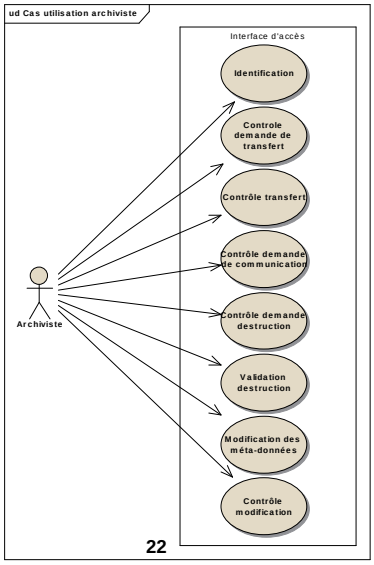
Les cas d'utilisation « Fonctionnel »



Les diagrammes d'activités



Les cas d'utilisation « Acteur »



Le transfert

- **UCC6** : un utilisateur du service versant ou un archiviste effectue un transfert
- **UCC8** : un archiviste contrôle un transfert
- **UCC9** : un utilisateur du service versant reçoit le résultat d'un transfert

Cas d'utilisation (1)

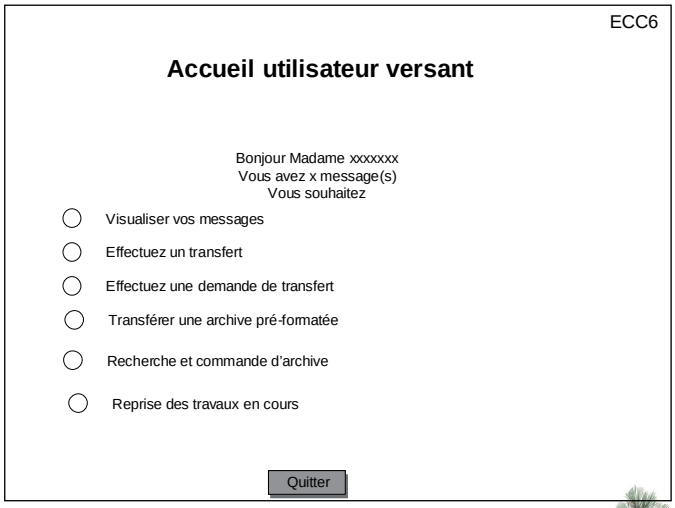
- **UCC6**
- **Objectif :** Un utilisateur du service versant ou un archiviste effectue un transfert

- **Scénario principal :**
- 1 Le système affiche un écran d'accueil (ECC6)
- 2 L'utilisateur sélectionne « effectuer un transfert »
- 3 Le système affiche un écran de saisie des informations de l'archive (ECC3) L'utilisateur saisit les informations de l'archive.
- 3a Dans le cas ou il y a eu une demande de transfert préalable, l'utilisateur saisit l'identifiant de la demande de transfert. Dans ce cas, les informations des métadonnées seront déjà renseignées et pourront être modifiées (Toutefois, dans le cas ou des métadonnées accompagnent le transfert, celles ci remplacent les métadonnées fournies lors de la demande de transfert).
- 4 Le système affiche un écran de structure de l'archive (ECC2)
- 5a L'utilisateur peut ajouter un objet
- 6a Le système affiche un écran de saisie des informations d'un objet (ECC4).Les champs remplis au niveau supérieur (archive ou objet) sont proposées par défaut aux niveaux inférieurs
- 7a L'utilisateur saisit les informations spécifiques de l'objet (différentes des informations du niveau supérieur)

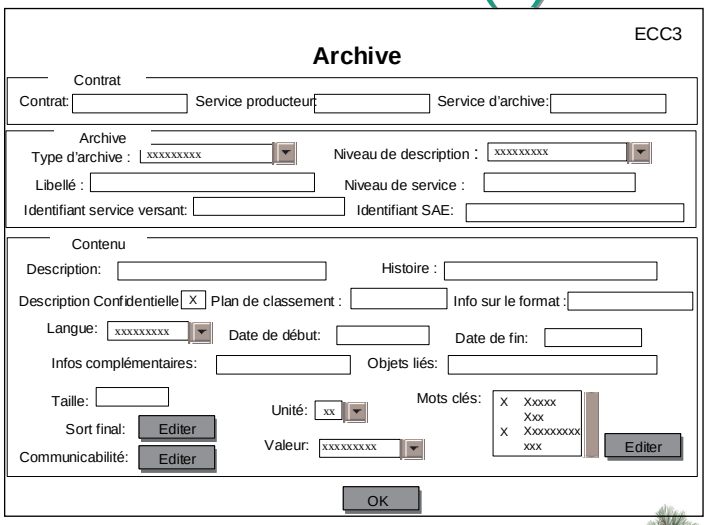
Cas d'utilisation (2)

- 8a Le système revient en 5
- 5b L'utilisateur peu ajouter un document
- 6b Le système affiche un écran de saisie des informations d'un document (ECC5)
- 7b l'utilisateur saisit les informations du document et charge le ou les fichiers joints
- 8b le système revient en 5
- 9 le système affiche l'écran de transfert (ECC15)
- 10 L'utilisateur saisit les informations de l'écran et valide le transfert.
- 11 le système revient en 1
- **Extensions :**
- **Variations :**
- 10 b l'utilisateur peut enregistrer l'état de la saisie et y revenir plus tard
- **Ecrans fonctionnels utilisés :** ECC2, ECC3, ECC4, ECC5, ECC6, ECC15

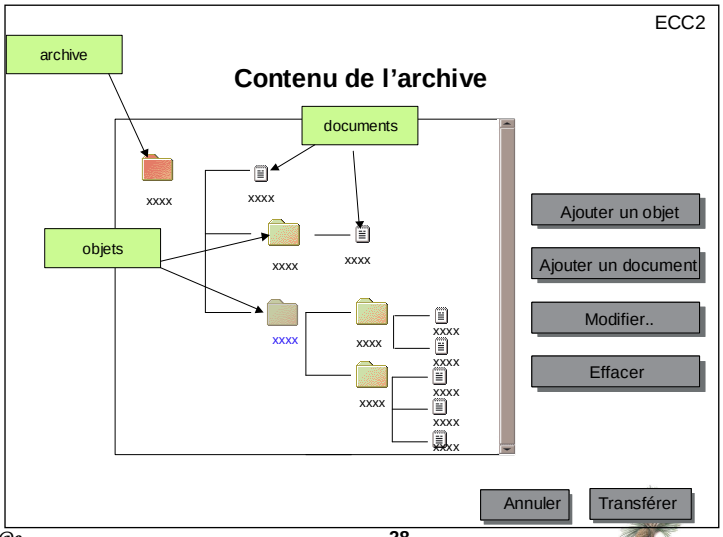
Écrans (1)



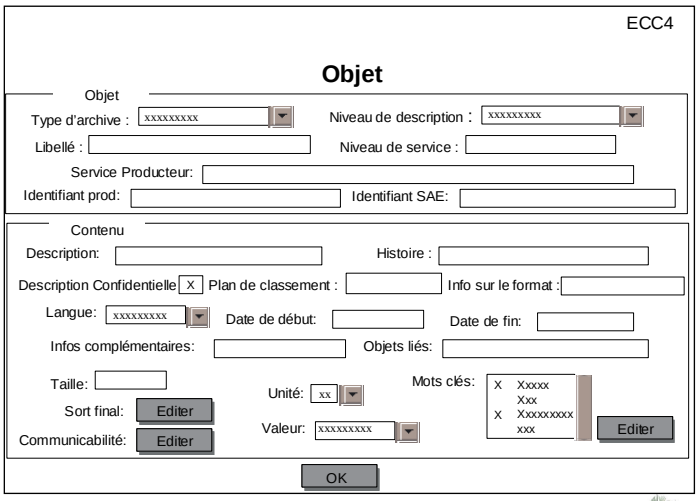
Écrans (2)



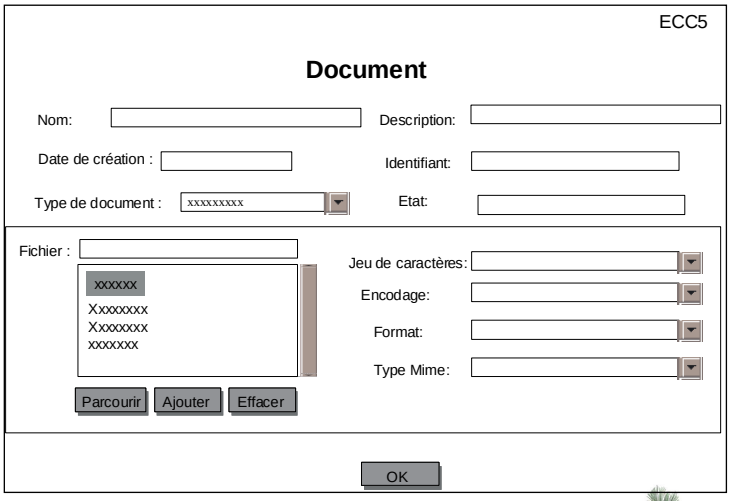
Écrans (3)



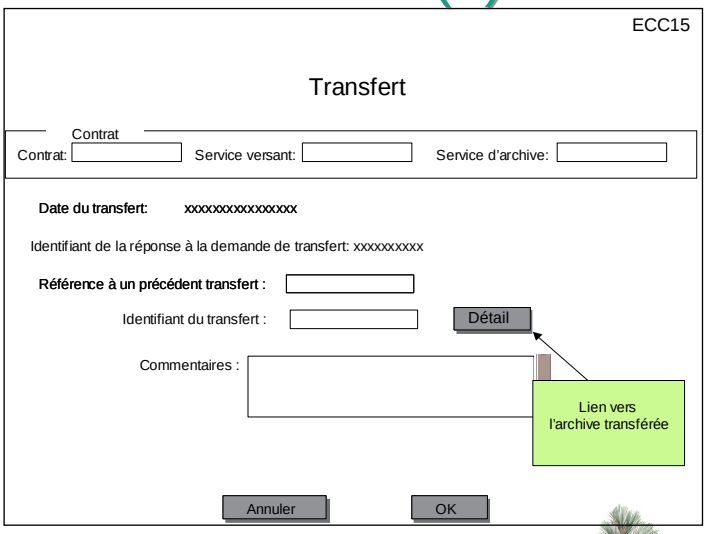
Écrans (4)



Écrans (5)



Écrans (6)



La structure « Contrat »

- Un contrat spécifie un engagement concernant les modalités de transfert d 'archives et d 'éliminations, ainsi que les modalités de communication entre un système producteur, un système versant et un système d'archivage
- Sur le plan technique:
  - Il définit les formats de documents acceptés
  - Il définit le type, et la structure des archives versées, ainsi que la périodicité des versements.
  - Il définit les utilisateurs habilités et leurs rôles
- La structure de données « Contrat » est utilisée par le système pour gérer les opérations et par les archivistes pour les contrôler

Principaux arbitrages : spécifications (1/2)

- Accepter les utilisateurs non authentifiés
  - gestion du grand public
- Prévoir des tables spécifiques aux contrats ; le fait que l'archiviste peut consulter les contrats
- Travailler sur les rôles et non sur les profils
- Distinguer deux rôles d'administrateur : administrateur fonctionnel et administrateur technique
- Prévoir deux identifiants pour l'archive : la cote du service versant et celle donnée par le service d'archive

Principaux arbitrages : spécifications (2/2)

- Établir une distinction claire entre les contrôles (manuels) et les vérifications (par le système)
- Donner le dernier mot à l'archiviste dans certains cas
  - Disparition du service producteur
  - En cas de vérification par le système que le versement n 'est pas acceptable (problème de format, de périodicité, de taille….), l 'archiviste est alerté et c'est lui qui prend la décision finale
- Introduire la gestion des registres des entrées
- Possibilité de modifier des versements une fois transférés dans l 'espace de stockage long terme

Outils de validation / conversion des formats et performances

- Formats en entrée
- Format cible d'archivage
- Outils de conversion
- Coût
- Temps de traitement

Les objectifs

- Définir les formats de fichiers destinés à l'archivage
- Définir les formats admissibles en entrée
- Sélectionner les outils de conversion
- Préparer les futurs choix
- Préparer les stratégies de conversions et les règles de gestion de celles-ci
- Connaître les outils disponibles et les coûts

Principaux arbitrages : formats et courriers électroniques

- Refuser les formats MS Office comme formats d'archivage
- Conserver les versions N, N-1 et la version initiale en cas de migration des formats
- Décomposer en entrée du système les différents éléments d'un courrier électronique (pièces attachées)
- Ne pas archiver des données compressées, chiffrées
- Ne pas détecter les doublons pour les PJ des courriers électroniques

Les formats cibles d'archivage

- Format très largement répandu
  - et/ou disposant d'une norme européenne ou internationale
- Les spécifications doivent être publiques et facilement accessibles
- La stabilité des formats doit être « raisonnable »
  - Une version nouvelle au maximum tous les 3 ans
- Il doit exister au moins
  - deux logiciels d'éditeurs différents disponibles sur le marché français ou européen
  - ou un logiciel en Open Source

Les logiciels de conversion

- Il faut rechercher en priorité des logiciels en Open Source
- Le moins de logiciels possibles
- Mode batch et mode interactif
- Qualité support technique
- Documentation utilisateur disponible et compréhensible
- Rapidité de conversion des logiciels
- Environnements conformes à l'environnement du pilote

Formats en entrée

Format image	PNG
	GIF
	JPEG 2000 (jp2)
	JPEG (jfif, jpg)
	TIFF
	BMP
	PCX
Foramat images animées	MPEG-2 Layer III (MP3)
	WAV
	MPEG-2
	MPEG-4
Format texte	HTML
	XHTML
	OpenDocument
	MSoffice DOC
	Msoffice XLS
	Msoffice PPT
	Msoffice DOC
	Msoffice XLS
	Msoffice PPT
	PDF/A
	PDF
	RTF
	CSV
	TXT

Format comprimé	ZIP
	WinRAR
	E-mail
Format structuré	XML
	XSD
	XML validé par XSD
	SGBD
Format plan	SVG
	DWG
	DXF
	CGM
	STEP

Formats cibles d'archivage

Images	*PNG*
	JPEG 2000
	TIFF
Image animée et son	*MPEG-2 Layer III (MP3)*
	MPEG-4
Document	*HTML*
	XHTML
	OpenDocument
	PDF/A (ISO 19005)
	CSV
	TXT
Format comprimé	*ZIP*
	WinRAR
	E-mail + pièces jointes
Format structuré	*XML*
	XSD
	XML
	TXT
Plan et format industriel	*SVG*
	CGM
	STEP

Outils de conversion

- Imagemagick
- Switch Sound File Conversion Software
- StarOffice 8 API Basic
- PDF Converter/4
- Batch ZIP ToolKit
- MSGDetach
- DWG to SVG converter MX

cas particulier
Développements spécifiques nécessaires pour les outils de mise en forme des bases de données à archiver

Les coûts

Imagemagick	Gratuit (Open source)
Switch Sound File Conversion Software	Gratuit ou 39$ pour la version professionnelle
StarOffice 8 API Basic	Gratuit (Open source)
PDF Converter/4	99 $ HT
Batch ZIP ToolKit	17 €/HT
MSGDetach	Shareware/ 10€HT
Outil spécifique pour chaque base	A définir
DWG to SVG converter MX	59,5 $/HT

Temps de traitement

Imagemagick	3 à 4 secondes par images (suivant type image)
[2]	
Switch Sound File Conversion Software	Pas d'information sur les performances
StarOffice 8 API Basic	5 à 30 secondes par document
PDF Converter/4	2 à 3 secondes par pages testés
Batch ZIP ToolKit	Variables suivant tailles fichiers
MSGDetach	1 à 2 secondes par mail au format eml
Outil spécifique pour chaque base	Suivant type de base
DWG to SVG converter MX	2 à 20 secondes par document suivant complexité du plan

Gestion des anomalies

Typologie des anomalies

- Anomalie bloquante
  - Le logiciel ne peut effectuer la conversion du format d'entrée vers le format cible correspondant
- Les anomalies non bloquantes mais détectables
  - Le logiciel de conversion indique lors du traitement qu'il existe une erreur ou plusieurs erreurs mais continue le traitement
- Les anomalies non détectables

Gestion des anomalies

- Si les anomalies portent sur un part limitée d'un versement, l'archiviste décide ou pas de rejeter le tout

Traçabibilité des opérations

- Identifiant de la machine ayant effectué le traitement
- Nom du logiciel de conversion et version
- Identification du traitement
- Nom du fichier à convertir
- Format d'entrée et en sortie
- Date et heure de la conversion
- Éventuellement anomalies

Architecture technique

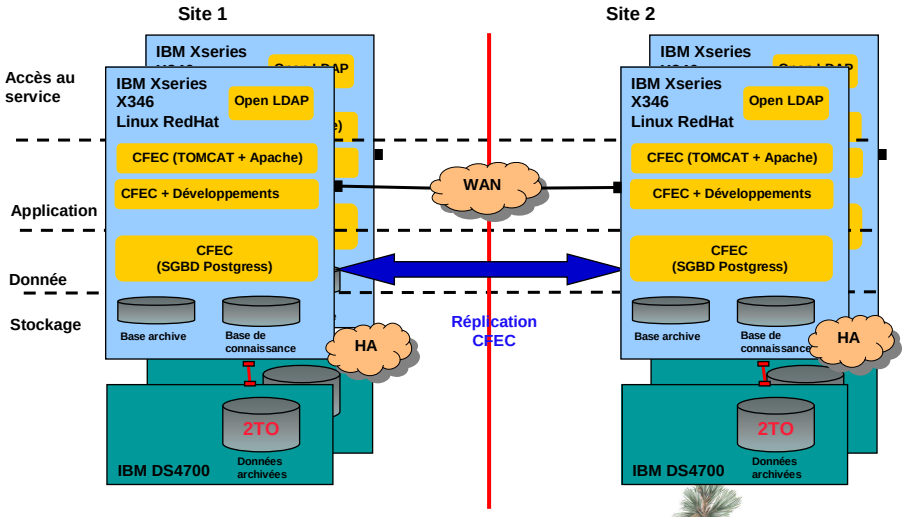


Tableau de synthèse des besoins opérationnels 1/2

Critère		Formulation de l'information - Statut
1/ Continuité de service		Actuellement disponibilité : jours et heures ouvrées (8H-20H) Même niveau de disponibilité pour le pilote.
2/ Recouvrement		Double réplication intra-site et inter-sites
	RPO	Temps de reprise maximum : ½ journée
	RTO	Transfert et incident avant fin réplication => Reprise du transfert
	Bascule arrière	Besoin d'un site de repli en cas d'incident majeur
3/ Emplacement		Deux sites distants de 60 km environ
4/ Protection système Intégrité données		Mesures de protection : 3 points de vue complémentaires • la protection au niveau système, • la protection au niveau des bases de données et d'annuaire, • la protection des objets.
5/ Sécurité Intégrité Cohérence		Sécurité: authentification et habilitation Conservation intègre espace de stockage Cohérence : données, objets documents

Tableau de synthèse des besoins opérationnels 2/2

Critère	Formulation de l'information - Statut
6/ Intégration dans le système d'information	Contraintes de deux sites •St Cyr •Fontainebleau
7/ Environnement de production	
Volumétrie T0/T1/T2	2 To (unitaire sur chaque serveur) . Evolution espace stockage reste à préciser =>T1,T2
Performances	Pas temps-réel critique Niveau de performances standard
8/ Conservation	
Durée de conservation	Définie dans les contrats et connue au moment du transfert vers le pilote
Gestion des médias	1 seul type de support dans le pilote Simulation de gestion de supports différents
9/ Exploitation	Cohérence avec exploitation à Saint Cyr Utilisation outil NAGIOS

La solution CFEC/IBM Continuité niveau système : redondance standard

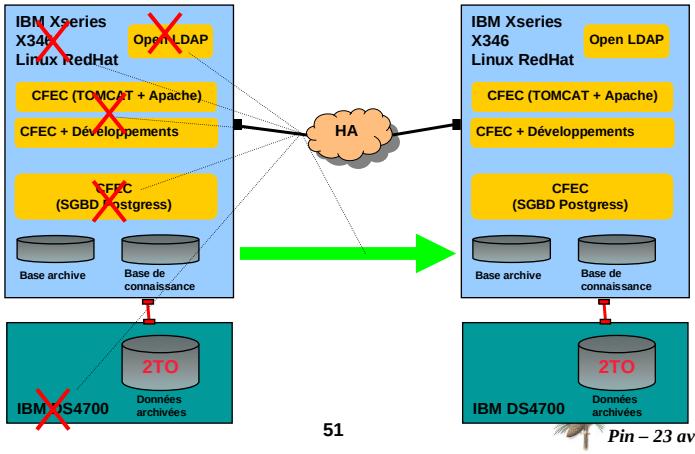
Ensemble de moyens de redondance au niveau matériel:

- Redondance composants du serveur (Xseries 346)
- Redondance disques (système) RAID1 (Xseries 346)
- Redondance disques données RAID5 (baie stockage DS4700)
- Redondance des serveurs
- Redondance des accès
- Redondance baies de stockage

Support / Intervention en cas de défaillance

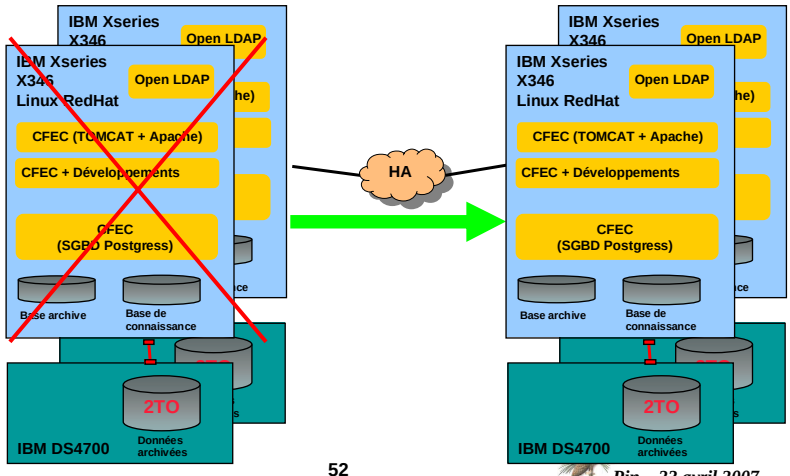
Réponse aux besoins de continuité

Continuité de la solution
Incident local mineur => pas de bascule (redondance matérielle)
incident local majeur => bascule locale



Réponse aux besoins de recouvrement

incident site 1 => bascule sur site 2



Site 1 Site 2

La solution CFEC/IBM
Réponse aux besoins de Sécurité

CRITERE	*Composants Solution*
Authentification	Annuaire LDAP + login/password (ou certificat électronique X509V3)
Contrôle d'accès	Rôle + Contrat + BDD
Intégrité de l'environnement (cloisonnement des services)	Séparation responsabilités (administration technique et administration fonctionnelle Limitation des droits
Intégrité des informations (protection / modifications non autorisées)	Empreinte intégrité SHA1+horodatage+scellement/signature électronique
Confidentialité	Non significatif pour le pilote
Disponibilité	Infrastructure opérationnelle
AUDIT : enregistrement des activités pour reconstitution transactions et/ou processus.	Module de traçabilité du coffre-fort électronique + Conservation des journaux dans le coffre-fort électronique
TRAÇABILITE : authentification de l'origine et du destinataire des transactions	Module de traçabilité du coffre-fort électronique
NON-REPUDIATION.	Scellement interne au CFEC par signature électronique Signature des messages : conservation + preuve de vérification