Bibliothèque nationale de France (BnF)

Système de Préservation
et d'Archivage Réparti (S.P.A.R.)

Plan

- Introduction
  - → Contexte et objectif
- L'infrastructure
- Démarche pour la réalisation
  - → Les groupes fonctionnels
  - → L'étude technique
  - → Le cahier des charges
  - → Description générale du système

Introduction

- 1997⇨2007 : 10 ans d'offre de services « numériques » à la BnF.
- Contribution et suivi au niveau international des projets de conservation des documents numériques n'ayant pas d'équivalents analogiques
- Projet SI numérique de la BnF
  - → Système homogène d'archivage, de conservation à long terme de l'ensemble des ressources numériques patrimoniales.
  - → Évolutions des applications existantes pour qu'elles utilisent, à terme, ce nouveau magasin virtuel.
  - → Projet d'une durée de quatre ans environ.

Le contexte

- Production documentaire uniquement sous forme numérique
  - → Dépôt légal : WEB, Presse, etc.
- Dématérialisation des supports analogiques
  - → Opérations de numérisation de masse des collections papiers
- Production numérique suite à l'obsolescence voire la disparition des équipements de restitution
  - → Numérisation de sauvegarde des collections audiovisuelles
- Disparition progressive des moyens de productions de microforme
  - → Remplacement des moyens de capture numérique

État des lieux

- Diversité de supports/formats d'enregistrement
  - CD-ROM, CD-R, CD-Century, Disque Optique, LTO, DLT, VHS, DVD, Disque, etc.
- Diversité de formats de fichier/de structure
  - TIFF, JEPG, MEPG, HTML, XML, etc.
- Diversité de durée de vie
- Diversité de procédures
  - DAV, DSI, DBN, DREP, Mission archives
- Diversité de moyens

Les problèmes posés

- Dépendances entre les données créées et l'environnement de création
- Obsolescence des technologies de stockage, des logiciels et des systèmes
- Absence de données ou de documents descriptifs
- Garantir l'accès au contenu sans modifier l'information et sa représentation
- Accès parmi des quantités exponentielles
- Adaptation à des techniques de recherche d'information de plus en plus performantes

Le projet SPAR

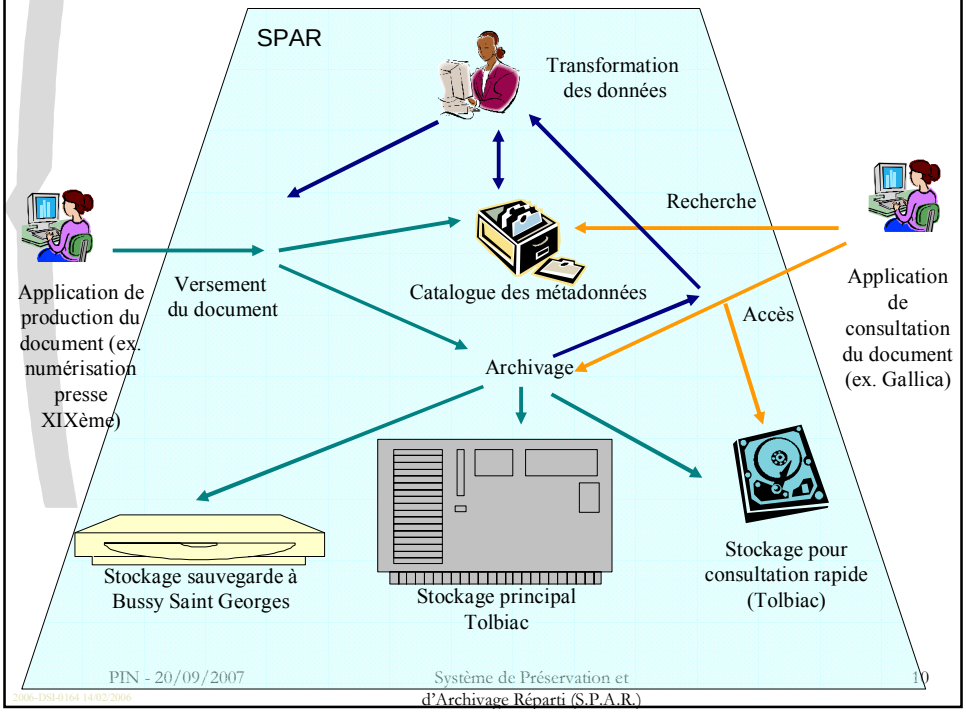
- Dispositif dont la vocation est de préserver l'information pour permettre à une Communauté Définie d'Utilisateurs d'y accéder et de l'utiliser.
- Imaginer et concevoir un système OAIS qui puisse permettre à la BnF de:
  - → PRESERVER son patrimoine numérique
  - → ARCHIVER l'ensemble de ses données
  - → REPARTIR l'accès à ses données
- Nom de code: SPAR
- Mise en œuvre du système en 2 temps :
  - → Stockage (marché infrastructure)
  - → Versement et gestion des données et accès (marché réalisation)

Historique du projet

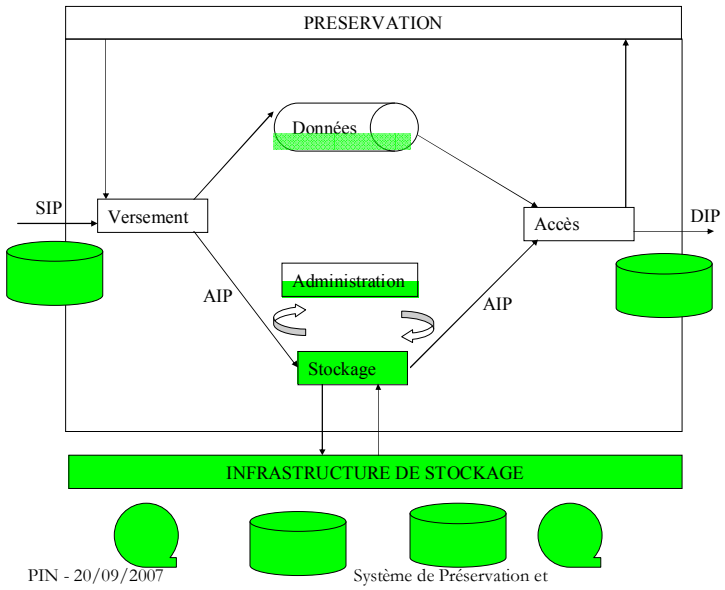
- 2005 : BnF décide d'avoir un système de préservation numérique
  - → Novembre 2005 : SUN-StorageTek remporte le marché d'infrastructure (SPAR-Infra)
- 2006 : BnF organise le SI-Numérique
  - → Mars 2006 : Définition de l'organisation et des groupes de travail
- 2007 : BnF définit un système complet
  - → Mai 2007 : AO pour la partie applicative (SPAR-Réalisation), remise des offres (le 24 septembre)

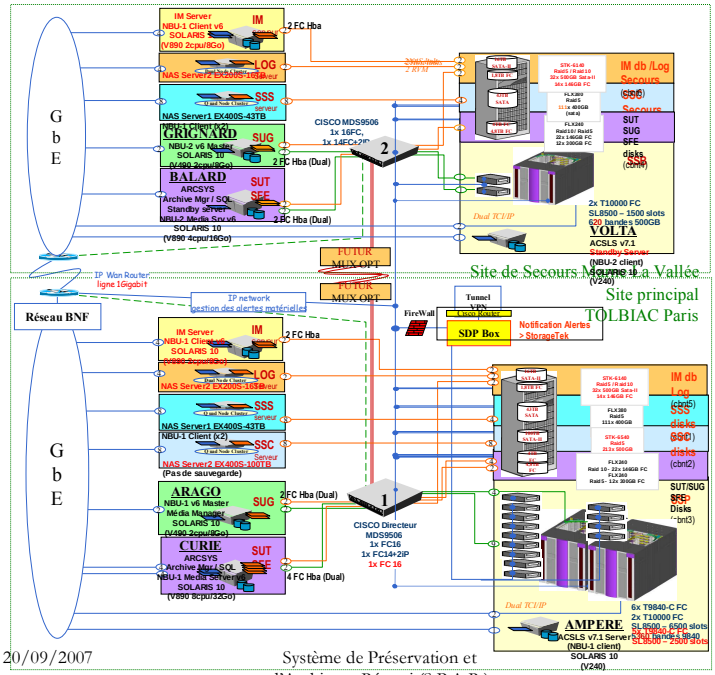
Archivage et Préservation : Objectif à la BnF

- Prendre la responsabilité de l'archivage de données (norme OAIS)
- Fournir un service de stockage d'archive
- Atteindre une masse critique minimale pour réduire les coûts (moyens matériels, logiciels et humains)
- Permettre la mutualisation de stockage d'archives entre plusieurs établissements ou institutions distinctes



OAIS / SPAR-Infrastructure





Stockage principal et secours

Sun StorageTek SL8500

- jusqu'à 64 lecteurs
- jusqu'à 8500 cartouches
- jusqu'à 8 bras
- jusqu'à 32 robotiques liées

Stockage principal
340 To par robotique
Stockage de secours
4,25 Po par robotique

Stockage secondaire

Sun Stk FLX240

- 111 disques de 400 Go
- RAID 5
- Disques SATA

- 33 To utile

Stockage de consultation

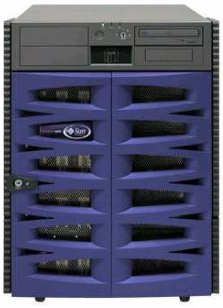
Sun Stk 6540

- 213 disques de 500 Go
- RAID 5
- Disques SATA-2

- 76 To utile

Serveurs de pilotage

Serveur de traitement
Sun Fire V890



8 Processeurs 1,5 GHz
32 Go mémoire
4 HBA dual FC
Solaris 10

Serveur de gestion
Sun Fire V490

4 Processeurs 1,5 GHz
16 Go mémoire
2 HBA dual FC
Solaris 10

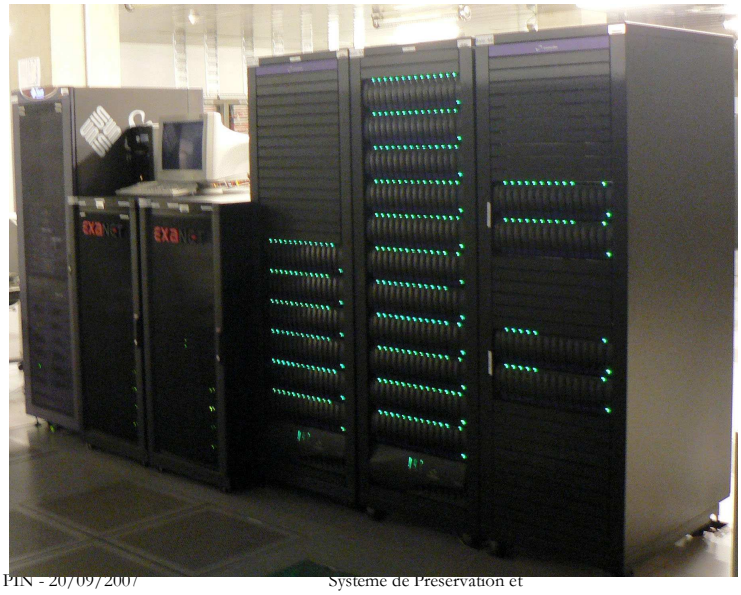
Salle informatique

Salle informatique

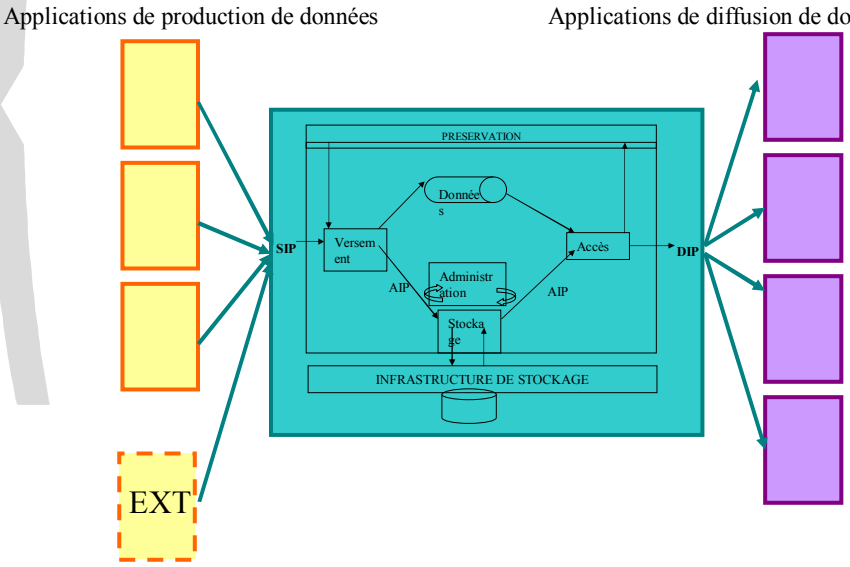
Salle informatique

Salle informatique

Salle informatique



SPAR Réalisation : contexte



SPAR Réalisation : démarche d'étude

- Définition du besoin (Groupes fonctionnels)
- Étude technique des solutions possibles
- Présentation au comité de pilotage
- Rédaction du cahier des charges de réalisation

SPAR Réalisation : groupes fonctionnels

- Aspect fonctionnel : définir les besoins du système
- 5 groupes fonctionnels :
  - → G1 : fonctions
  - → G2 : communauté d'utilisateurs
  - → G3: modèles d'information (définition des métadonnées et des paquets d'information)
  - → G4 : gestion des risques
  - → G2D : gestion des droits

Les producteurs de données (1)

- Département de la conservation, Service numérisation
  - → Numérisation interne : plans de préservation des originaux ;
  - → Marchés de numérisation : sélections pour la bibliothèque numérique Gallica ;
- Département de la reproduction
  - → Demandes de clients ;
- Département de l'Audiovisuel :
  - → Conservation : plans de sauvegarde de documents originaux audio et vidéo ;
  - → Dépôt légal des phonogrammes et des vidéogrammes;
  - → Dépôt légal multimédia multisupport;
  - → Dépôt légal des documents électroniques :logiciels, bases de données, jeux ;

Les producteurs de données (2)

- Département de la bibliothèque numérique
  - → Dépôt légal du web : sites sélectionnés pour la collecte ;
  - → Dépôt volontaire de publications électroniques (PQR, élections…) ;
- Mission des archives
  - → Archives administratives de la BnF ;
- Délégation à la communication
  - → Publications de communication de la BnF ;
- Délégation à la diffusion culturelle
  - → Publication de la BnF en ligne ou sur papier : expositions, dossiers pédagogiques, éditions de la BnF ;
- Direction des collections (DCO)
  - → Acquisition sous forme électronique : CD-ROM, publication au format PDF.

Les utilisateurs (1)

- Bibliothèque numérique pour le web (Gallica)
  - → Documents de la bibliothèque numérique à partir d'Internet ;
- Bibliothèque numérique en interne (Renet)
  - → Documents de la bibliothèque numérique et des CD-ROM en Intranet ;
- Postes consultation audiovisuelle (PAV)
  - → Documents du département de l'Audiovisuel en salles de lecture accessibles aux chercheurs accrédités ;
- Département de la reproduction
  - → Documents pour le service client (fourniture de copies numériques, transformées ou non, en ligne ou sur CD-ROM, impression à la demande) ;

Les utilisateurs (2)

- Délégation à la communication
  - → Documents en vue de publications sur le web institutionnel et de catalogues, brochures, etc. ;
- Délégation à la diffusion culturelle
  - → Documents en vue de publications éditoriales ou en ligne sur le web ;
- Service archive
  - → Documents produits par les agents sur les postes en intranet ou en ligne selon les utilisateurs autorisés.

Volumétrie

		2005	2006	2007	2008	2009	2010	2014
		Période	*Période*	*Période*	*Période*	*Période*	*Période*	*Période*
		1	2	3	4	5	6	7
Entité/ projet	**Formats**							
SA/ conversion Video, doc sonores	Mpeg-2, wav	0	0	0	0	400	412	460
SA/ CD audio pressés à migrer	CD-audio	0	0	0	0	50	56	80
SA/ vidéo à migrer								200
DBN/ Archives Web	html, PDF, Word, JPEG , GIF, PNG, fichiers AVI, RealAudio, RealVideo, MPEG audio et video, Java applets, fichiers Flash ,,,	80	160	240	320	400	480	800
DSC/NUM Marchés en cours	JPEG, TIFF	1,6	1,7	0,9	1	1	1,05	1,25
DSC/NUM Dunhuang	TIFF	1	1	1	1	1	1	1
DSC/NUM interne presse	TIFF	6,4	12,8	19,2	25,6	32	38,4	64
DSC/NUM externe presse	TIFF	9,5	19	28,5	38	47,5	57	95
DSC/NUM préservation	TIFF	1	2	3	4	5	6	10
DSR/ DRE	TIFF	3,07	3,77	4,47	5,17	5,87	6,57	9,37
PQR/ Collecte électronique	PDF	1,82	3,62	5,42	7,22	9,02	10,82	18,02
TOTAL en To		**104**	**204**	**302**	**402**	**951**	**1 069**	**1 739**

Définition des filières

- Fondées sur les relations des données numériques / numérisées et du système d'archivage numérique.
- Envisagées sur les exigences attendues / possibles de part et d'autre et, ce, aux différents moments du dialogue : entrée, maintenance, accès.
  - → Du point de vue système d'archivage vis-à-vis du producteur.
    - ❖ Négociation préalable (possible ou non)
    - ❖ Cadre législatif et réglementaire (impératif ou non)
    - ❖ Conditions d'accès
  - → Du point de vue producteur vis-à-vis du système d'archivage
    - ❖ Fixité des données (modification, version, élimination ou non)
    - ❖ Pérennité des données (données, métadonnées, structures et représentations)
    - ❖ Accessibilité (directe / applicative, immédiate / différée)

Les filières

- Numérisation de conservation
- Numérisation de reproduction
- DL automatique (support, Web de surface)
- DL négocié (PQR, Web profond)
- Productions administrative et technique
- Dépôt / Tiers archivage
- Acquisition / Don

Les filières : cadres et exigences

Filière	Caractère de négociation	Exigence de conservation
Numérisation de conservation	Cadre contractuel	Fixité Pérennité
Numérisation de reproduction	Cadre contractuel	Pérennité Remplacement possible
Dépôt légal automatique	Cadre législatif	Fixité Pérennité
Dépôt légal négocié	Cadre contractuel	Pérennité
Productions administrative et technique	Cadre législatif / contractuel	Fixité Limitation dans le temps
Dépôt / Tiers archivage	Cadre contractuel	Fixité Limitation dans le temps
Acquisition / Don	Cadre contractuel	Pérennité

Les politiques (1)

- Politique de versement : caractéristiques de la négociation et du protocole de versement
  - → Qui négocie ?
  - → Que négocie-t-on ? : droit, format, volumétrie, flux
- Politique d'archivage : caractéristiques de la conservation
  - → Que conserve-t-on ? : données, métadonnées, systèmes de représentation… original, master, produits dérivés
  - → Comment conserve-t-on ? : à l'identique (train original des bits), émulation, migration
  - → Combien de temps conserve-t-on ?

Les politiques (2)

- Politique d'accès : caractéristiques de l'accès
  - → Avec ou sans restriction
  - → Avec ou sans services supplémentaires (ex. veille des formats)
  - → Immédiat ou différé
  - → Direct (applicatif DSI) ou indirect
  - → Sur place, à distance
  - → Volumétrie des transactions

SPAR Réalisation : étude technique

- Groupe transverse DSI d'étude technique
- Mise en place de 2 évaluations
  - → Évaluation à « grosse maille » : large nombre de solutions / nombre limité de critères
  - → Evaluation à « petite maille » : évaluation vis-à-vis des contraintes BnF des solutions préalablement sélectionnées

=> élaboration de 3 scénarios au plus

Evaluation « grosse maille »

ADORe (IR)	DSpace (IR)	LOCKSS (IR)
ARCSys (AM)	e-prints.org (IR)	myCORE(IR)
ARNO (IR)	FEDORA (IR)	Octopus (AM)
BePress (IR)	FileNet (IR)	OpenText (CM)
Cdsware (IR)	GreenStone (IR)	STAR (AM)
DIAS (AM)	Hummingbird (CM)	SaTStore (AM)
Documentum (CM)	Interwowen (CM)	

IR : Institutional Repository, AM : Archive Manager, CM : content manager

Evaluation « grosse maille »

Analyse selon 3 axes

Fonctionnel

- Peut intégrer les identifiants pérenne ARK?
- Peut intégrer un système d'authentification ?
- Gère des métadonnées description de manière extensible ?
- Gère des métadonnées de pérennisation liés aux objets de données/documents ?
- Comprend des fonctions d'entrées ?
- Comprend des fonctions d'accès ?
- Comprend des fonctions d'administration ?
- Gère différents types de format ?
- Gère le cycle de vie des documents (workflow) ?
- Comprend des fonctionnalités de migration (gestion, suivi, etc.) ?

Qualité technique

- Semble extensible ?
- Est intéropérable avec un système de stockage ?
- Est intéropérable avec un système de supervision (SNMP, etc.) ?
- Est intéropérable avec un annuaire (LDAP, X509 etc.) ?
- A des références d'implémentation de grande taille (équivalent à la BnF) ?

Pérennité

A pour objectif d'implémenter le modèle OAIS ?

- Est modulaire ?
- Est libre/ouvert ?
- Est maintenable (organisme de maintenance, qualité du support, communauté, documentation)
- Est mature (références nombreuses, longue expérience, etc.)

Evaluation « grosse maille »

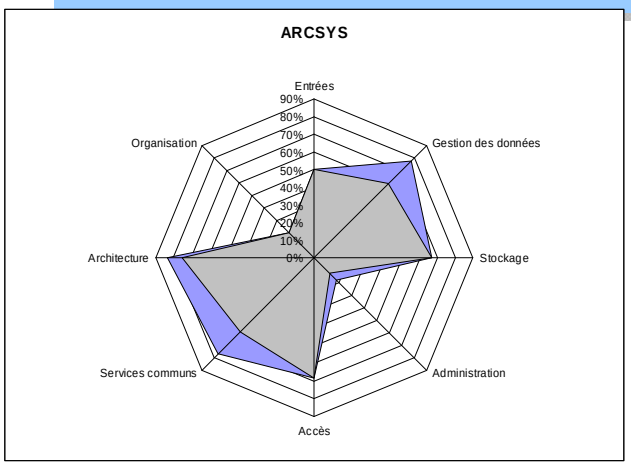
	Catégorie	Fonctionnel	Pérennité	Technique	Pondération
					3*F+3*P+T
aDORe	Institutionnal Repository	14	5	4	**61**
ARCSys	Archive OAIS Manager	18	4	7	**73**
ARNO	Institutionnal Repository	10	1	2	**35**
bepress/DigitalCommons	Institutionnal Repository	4	4	0	**24**
CDSware	Institutionnal Repository	15	7	4	**70**
DIAS	Archive OAIS Manager	18	7	10	**85**
Documentum	Enterprise Content Management	13	5	4	**58**
DSpace	Institutionnal Repository	17	9	9	**87**
E-Prints.org	Institutionnal Repository	11	5	6	**54**
FEDORA	Institutionnal Repository	20	10	6	**96**
Greenstone	Institutionnal Repository	15	6	3	**66**
Hummingbird DM	Document Management	14	4	2	**56**
Intervowen RM	Record Management	16	4	2	**62**
LOCKSS	Institutionnal Repository	8	6	2	**44**
MyCoRe	Institutionnal Repository	15	6	4	**67**
Octopus	Archive OAIS Manager	15	5	4	**64**
OpenText Archiving	Document Management	10	4	3	**45**
STAR	Archive OAIS Manager	14	3	4	**55**
saTSTORE	Archive OAIS Manager	16	8	4	**76**

Evaluation approfondie

- Solutions évaluées
  - → ARCSYS - DIAS
  - → DSPACE - FEDORA
  - → SatSTORE - STAF-R
- Grille en 132 critères répartis en 9 catégories

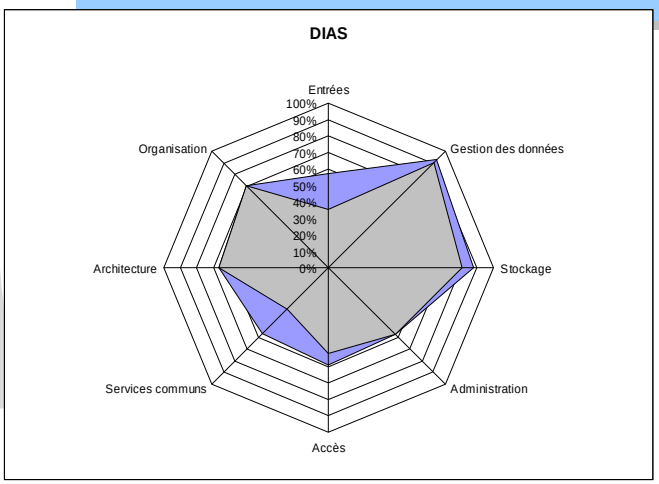
1. L'entité « Entrée »
2. L'entité « Gestion de données »
3. L'entité « Stockage »
4. L'entité « Administration »
5. *L'entité « Planification de la pérennisation »*
6. L'entité « Accès »
7. Les services communs
8. L'architecture
9. L'organisation

ARCSYS



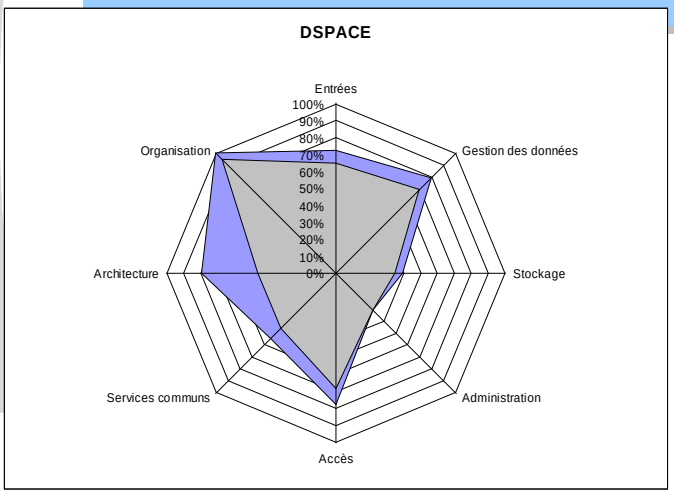
Min : 135 (55%)
Max : 153 (58%)

DIAS



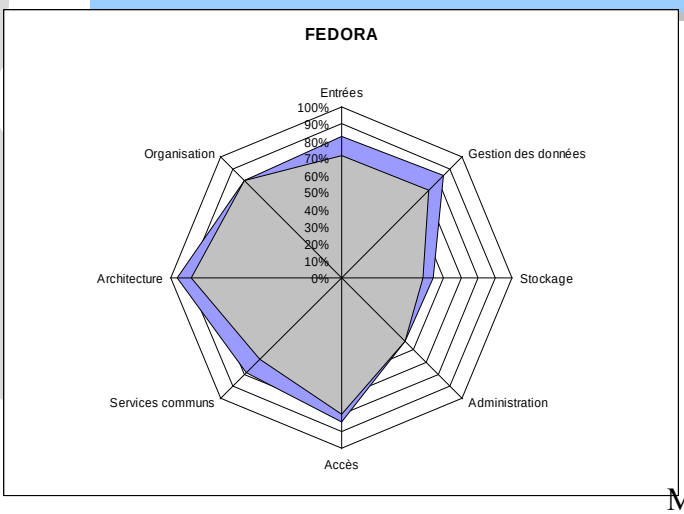
Min : 152 (58%)
Max : 179 (68%)

DSPACE



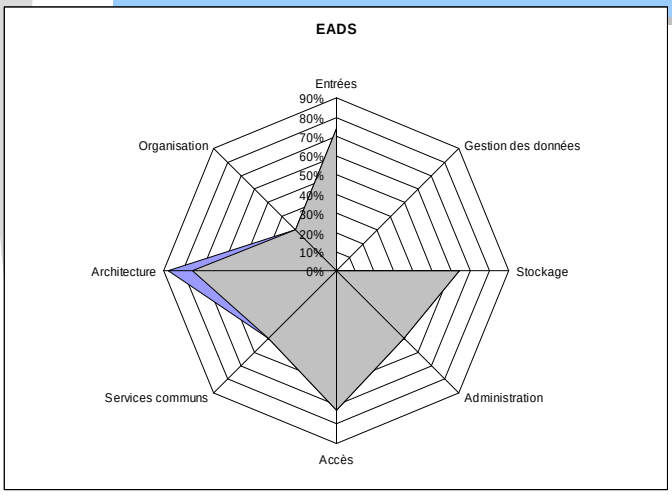
Min : 138 (52%)
Max : 158 (60%)

FEDORA



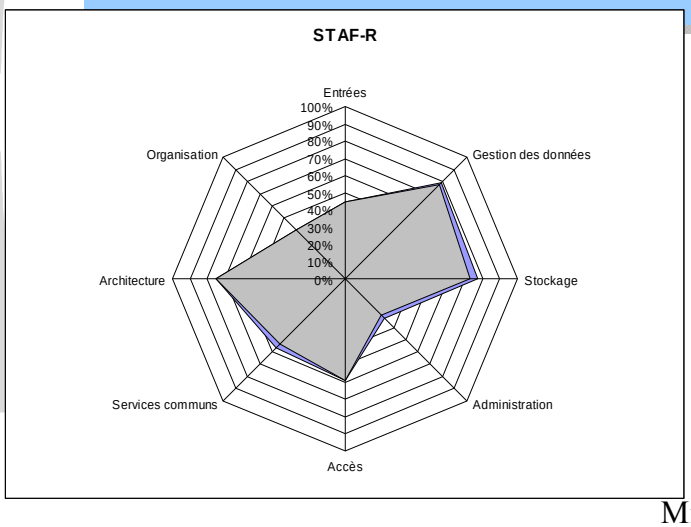
Min : 173 (66%)
Max : 193 (73%)

SatStore (EADS)



Min : 139 (53%)
Max : 141 (54%)

STAF-R



Min : 145 (55%)
Max : 150 (57%)

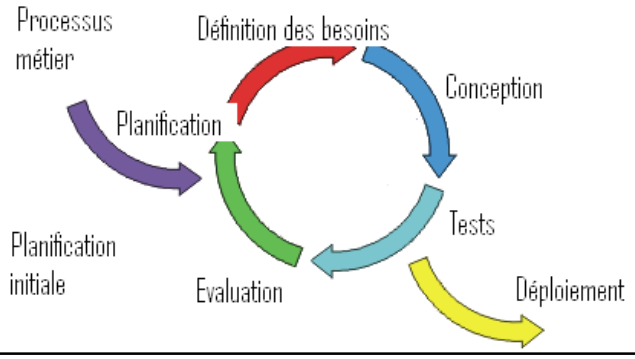
Scénarios possibles

- Aucune solution ne satisfait entièrement la cible …
- 2 scénarios envisageables :

1. Marché d'acquisition d'une solution de base s'appuyant sur un intégrateur pour acquérir/ajouter les briques manquantes
2. Marché de réalisation s'appuyant sur la solution Open Source Fedora + une interface avec le stockage (Arcsys ou autre)

SPAR : la réalisation du système

- Dans tous les cas, réalisation de manière itérative pour :
  - → prendre en compte l'aspect innovant du système,
  - → assurer au mieux l'acquisition progressive du système par les utilisateurs
- Étapes successives basées sur les filières pour couvrir de manière incrémentale l'ensemble des entités OAIS



SPAR : rédaction du CCTP

Entrants :

Rapport d'analyse des solutions
Documents issus du groupe de travail
Marché d'infrastructure

Définition des filières :

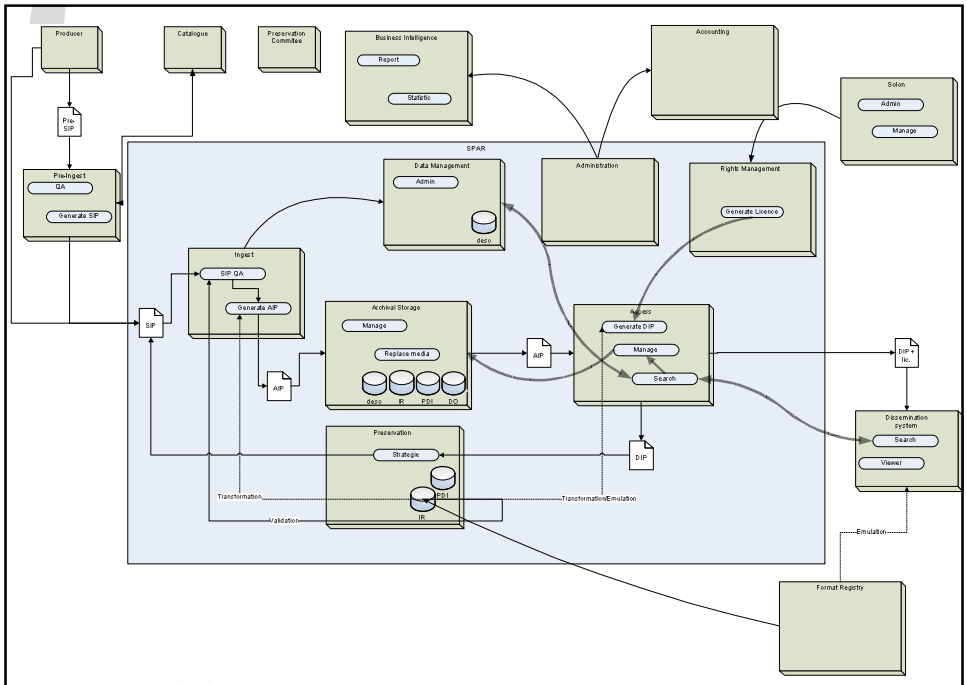
filière de numérisation de conservation,
filière de numérisation de reproduction,
dépôt légal des données audiovisuels,
dépôt légal des données électroniques,
archivage des documents administratifs,
tiers archivage

Définition du cahier des charges (1)

- A basée sur le modèle d'analyse RM-ODP (ISO 10746) : approche par perspectives
- Trois premières perspectives (entreprise, informationnelle et informatique) utilisées pour décrire le besoin
- Perspectives supplémentaires (technologie et ingénierie) laissées à l'appréciation du titulaire

Définition du cahier des charges (2)

- Modèle conceptuel général sur lequel s'appuie le système SPAR.
- Perspective entreprise
  - → Exprime formellement les besoins sous forme de cas d'utilisation.
- Perspective informationnelle
  - → Définit les types d'information au sein du système
- Perspective informatique
  - → Décomposition fonctionnelle du système en objets qui interagissent à travers des interfaces spécifiées de manière à faciliter la répartition



SPAR-Réalisation / OAIS