Projet pilote TGE-Adonis de mutualisation de l'archivage au CINES

CINES (O.Rouchon)

Groupe PIN – 13 Janvier 2009

Sommaire

- Le contexte du projet pilote
- La réutilisation de l'existant sur la plate-forme d'archivage du CINES
- Les développements spécifiques au projet pilote
- L'état des lieux
- Les perspectives

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 :

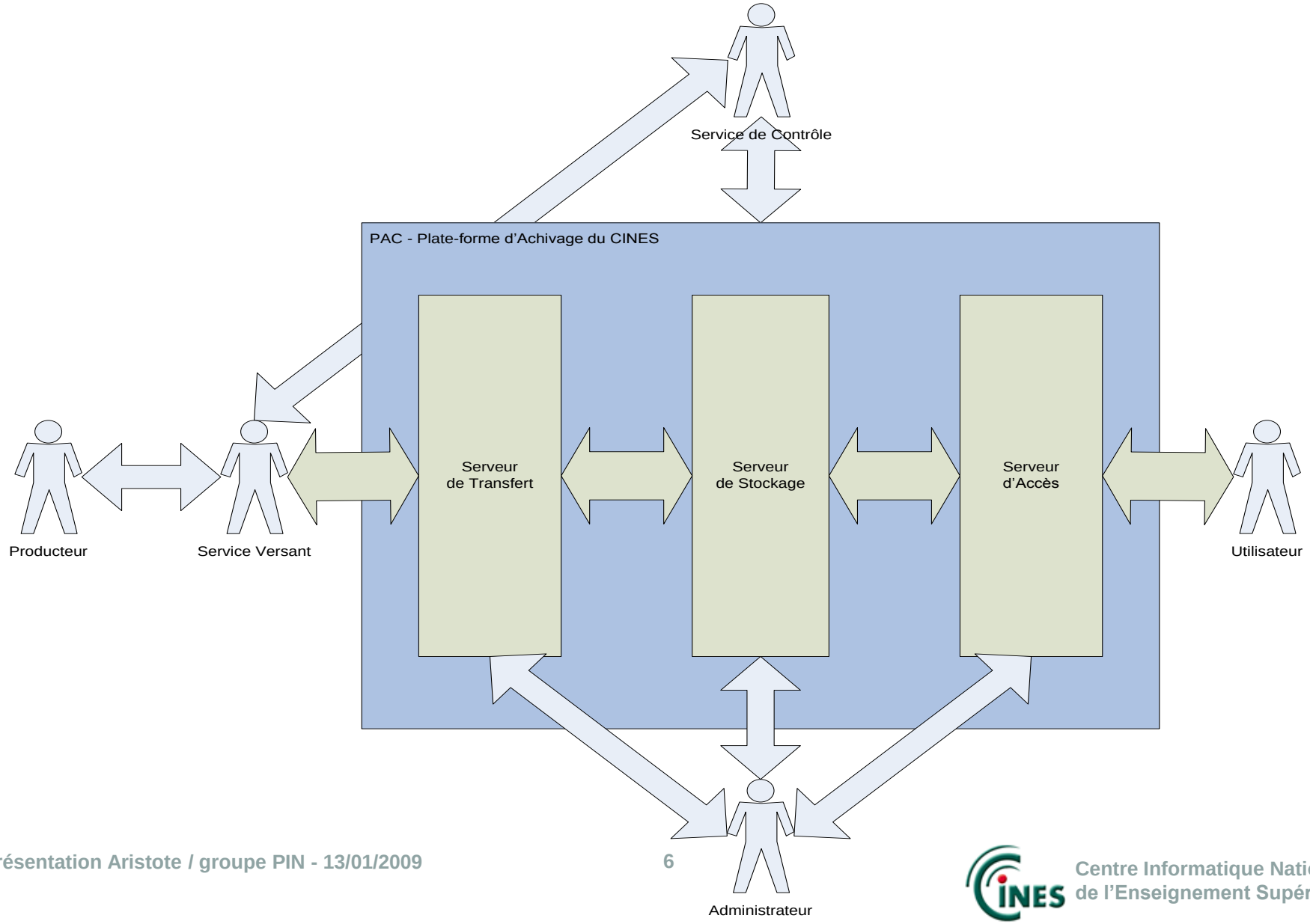
- Arrêté du 7 août 2006 relatif aux modalités de dépôt, de signalement, de reproduction, de diffusion et de conservation des thèses ou des travaux présentés en soutenance en vue d'un doctorat
- Convention du 2 mai 2007 (faisant suite à celle du 15 octobre 2003) relative à la mise en ligne et l'archivage 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 et l'archivage pérenne

La plate-forme PAC

PAC - Plateforme d'archivage au CINES - pérenne 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 volontairement limitée
  - Formats publiés, largement utilisés, normalisés si possible
  - HTML, PDF, TXT, XML, GIF, JPEG, TIFF, PNG, SVG, WAV
- Architecture basée sur du matériel SUN, le logiciel Arcsys et des logiciels libres
  - Java, MySQL, Jhove, ImageMagick
- Tous les projets d'archives partagent le même environnement
  - Mutualisation de l'infrastructure matérielle d'archivage
  - Protocole de versement et SIP génériques
  - Diminution des coûts de mise en place et d'exploitation
- Début de l'exploitation en production Mai 2008 – après migration de PAC v1.0

L’architecture logique de la plateforme



Les principes de fonctionnement

Transfert *(Entrées)*	**Réception des SIP**	*Détection d'un nouveau transfert* *Envoi d'un accusé de réception*
	Contrôle des SIP	*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*
	Création des AIP	*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 projets en cours

1. Deux projets pilotes en exploitation
   - Archivage des thèses électroniques
     - Documents nativement au format électronique versés par l'ABES
     - Fait suite à l'arrêté du 7 Août 2006
   - Archivage des revues SHS du portail Persée
     - Documents issus de la numérisation de revues au format papier dans le cadre du programme Persée
2. Deux projets en cours de réalisation
   - Archivage de documents sonores issus de la recherche dans le domaine de l'oral
     - Projet pilote CRDO dans le cadre du programme SHS du TGE-Adonis
   - Archivage de cours universitaires de Canal-U
     - Documents vidéos produits par le CERIMES
3. Un projet à l'étude
   - Archivage des documents déposés dans les archives ouvertes
     - HAL – Hyper Article en Ligne du CCSD

Le projet pilote CRDO

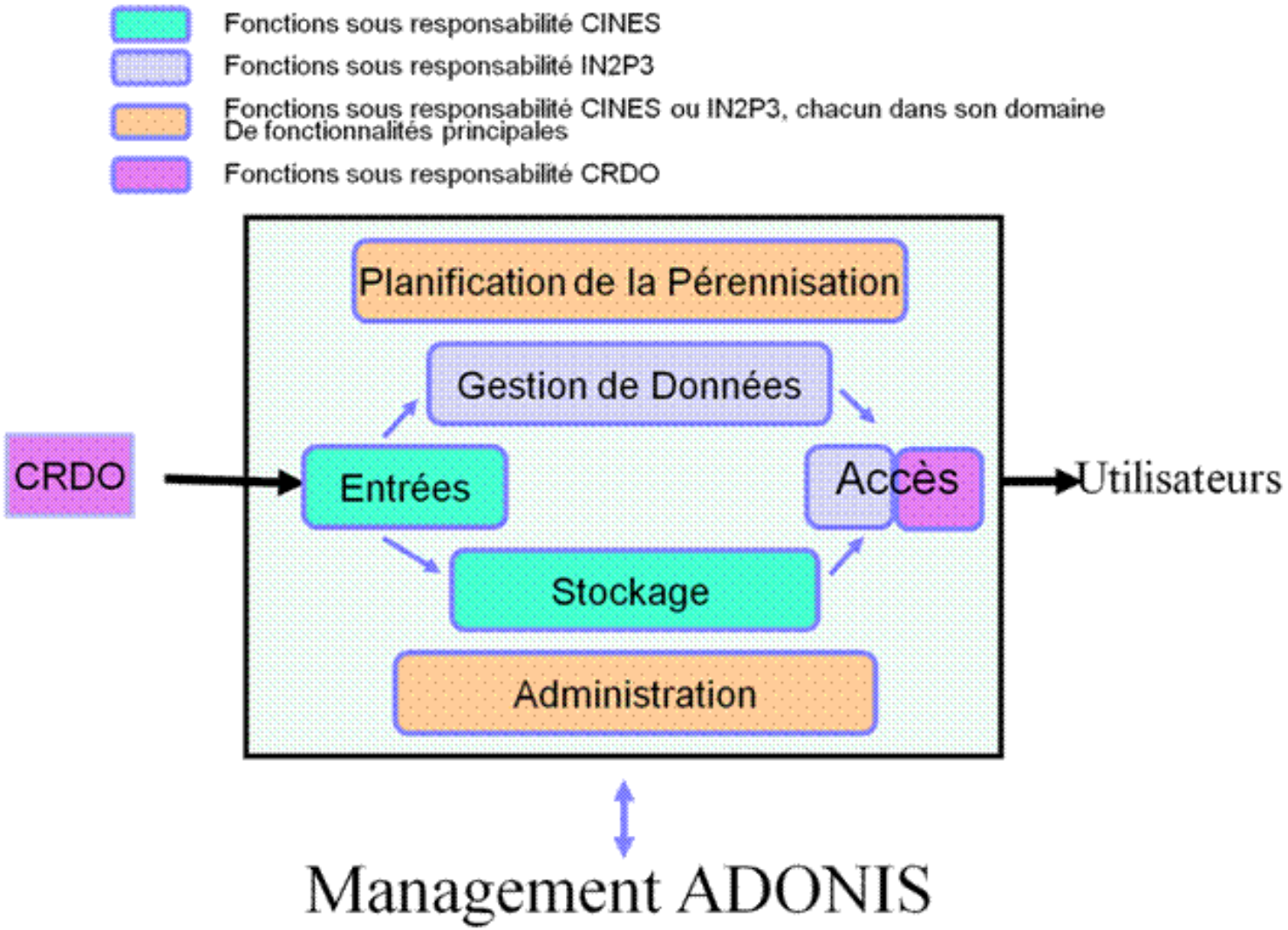
Le projet pilote – dont le TGE ADONIS est le maitre d'ouvrage – s'articule sur trois acteurs essentiels :

- Le CRDO, Centre de Ressources pour la Description de l'Oral, constitué
  - d'un groupe parisien - http://crdo.risc.cnrs.fr/exist/crdo/
  - d'un pôle à Aix-en Provence - http://crdo.up.univ-aix.fr/index.php?langue=fr
- Le CINES, Centre Informatique National de l'Enseignement Supérieur à Montpellier - http://www.cines.fr/,
- Le Centre de Calcul de l'IN2P3, Institut national de Physique Nucléaire et de Physique des Particules - http://cc.in2p3.fr/cc_accueil.php3?lang=fr.

L'objectif est la mise en place d'une infrastructure mutualisée pour la préservation et la diffusion de données de Sciences Humaines et Sociales – l'expérimentation porte sur des données « orales »

- Elles ne se réduisent pas à des enregistrements sonores
- Elles associent du texte (transcriptions, translitérations, annotations) et parfois des enregistrements vidéos
- Elles permettent donc de préfigurer partiellement la mise en place d'autres filières

La répartition des responsabilités

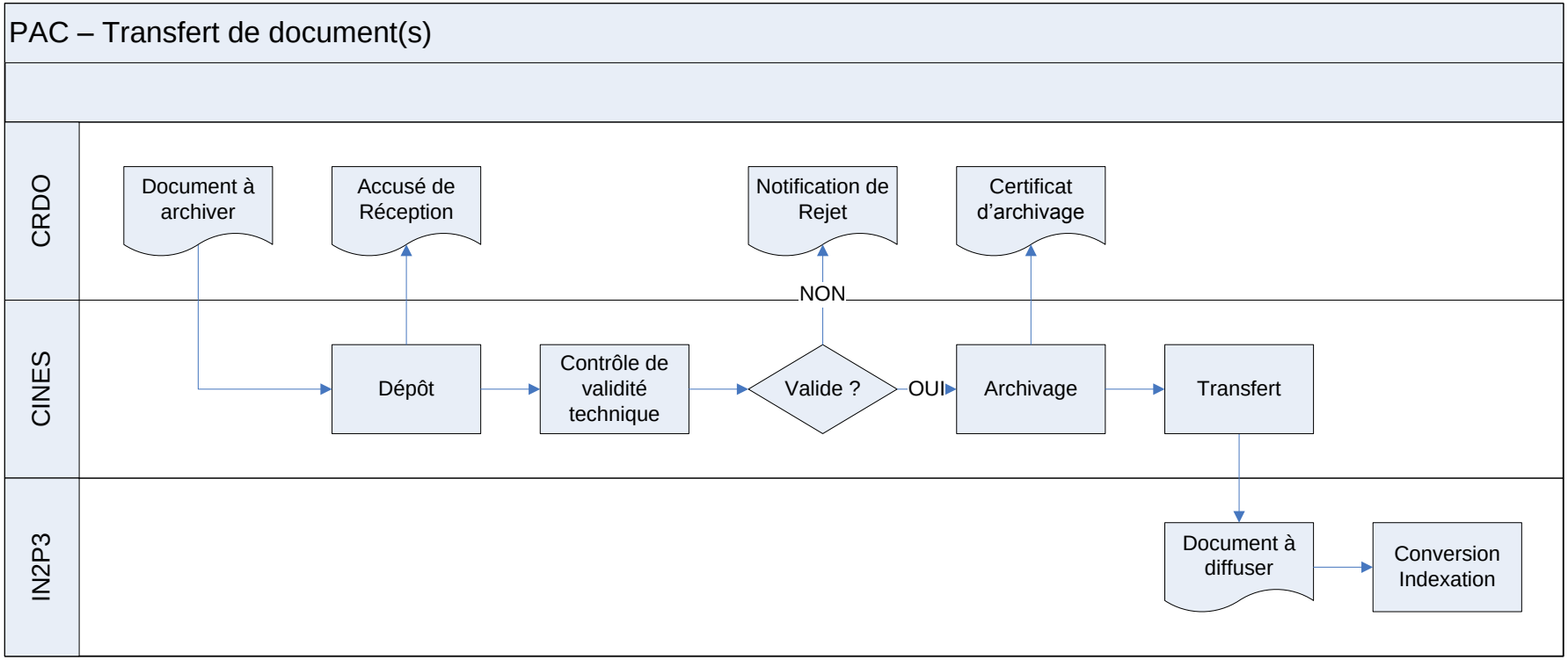


Le principe de fonctionnement

L'objectif est de réutiliser l'infrastructure en place au CINES, minimiser les développements spécifiques

- Dépôt des documents à archiver sur la plateforme du CINES pour archivage
  - Réutilisation du protocole de versement existant
  - Réutilisation de la structure du SIP existante
- Envoi des documents validés et archivés à l'IN2P3 pour diffusion
  - Prise en charge de nouveaux formats audio et vidéo,
  - Ajout d'un module de transfert du CINES vers l'IN2P3 – mode « push »
- Accès par la communauté d'utilisateurs aux documents via la plateforme de l'IN2P3

Les étapes du versement d'archives



Les formats de documents pris en charge

Formats identifiés et vérifiables :

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

Type	Format
Texte	HTML, PDF, TXT, XML
Image	GIF, JPEG, TIFF, PNG, SVG
Son	WAV, AIFF, OGG, AAC
Vidéo	mJPEG2000, MPEG4, OGG

JHOVE

ImageMagick

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

- Identifier,
- Valider,
- Caractériser,

Le format des fichiers transférés

Les besoins spécifiques du projet pilote

Outre la prise en charge de nouveaux formats de fichiers et la modification de la transaction de versement, trois besoins spécifiques exprimés par le CRDO ont été identifiés :

1. Possibilité de lier un ou plusieurs objets enrichissement/annotation à un enregistrement d'origine par une relation de parenté,

   Ajout de métadonnées génériques pour lier les SIP entre eux

2. Possibilité de mettre à jour certaines informations sans retransférer l'objet complet – c'est-à-dire verser les métadonnées descriptives (au format Dublin Core/OLAC) sans avoir à verser à nouveau l'objet lui-même s'il n'a pas été modifié,

   Mise en place d'une transaction de mise à jour de SIP

3. Possibilité de transmettre à l'IN2P3 des informations destinées à la diffusion - i.e. qui n'ont pas vocation à être préservées dans PAC

   Ajout d'un répertoire identifié – ignoré par le processus d'archivage et transmis à l'IN2P3

L’état des lieux en Janvier 2009

- Etude préliminaire en cours
  - Identification des objets à archiver terminée
  - Spécification des interfaces CRDO/CINES, CINES/IN2P3, CRDO/IN2P3 en cours
  - Analyse de l’impact sur les plateforme d’archivage, et de diffusion
  - Estimation des volumétrie à venir
- Développement et tests à venir
  - Interface de versement automatisé CRDO
  - Développements spécifiques sur la plate-forme PAC – prise en charge de nouveaux formats de fichier
  - Interface d’indexation et de conversion IN2P3
  - Plate-forme d’accès Fedora IN2P3
- Premiers tests d’intégration et utilisateur prévus en Septembre 2009
- Démarrage de l’exploitation en production fin 2009

Les perspectives

Ce projet pilote doit permettre de disposer de premiers éléments sur les coûts de fonctionnement de la solution, et de valider :

- Les fonctionnalités d'ensemble de la solution,
- Son caractère générique en dissociant ce qui est totalement applicable aux autres données des SHS de ce qui est spécifique des corpus oraux,
- La répartition des tâches et des responsabilités entre les acteurs,
- L'infrastructure matérielle-logicielle mise en place, sur le plan des performances et de la fiabilité,
- Les services permettant la recherche, la sélection, la récupération de données, avec la communauté des utilisateurs de corpus oraux,

A l'issue du projet pilote, un comité scientifique du TGE-Adonis évaluera les résultats produits

- Si le bilan est positif, une extension à d'autres données pourra être décidée,
- L'utilisation de la solution à d'autres Centres de Ressources Numériques sera alors généralisée.