

Mutualisation de la pérennisation et de l'accès aux données en SHS

Rapport d'avancement au 02/04/09 du projet pilote sur les données orales

Novembre 2008- Avril 2009

Rédaction : Claude Huc (consultant pour le TGE Adonis)

Contributions : CRDO, CINES, CC-IN2P3

Validation : TGE Adonis

Sommaire

1. Objet.....	4
2. Rappel du schéma fonctionnel du projet pilote.....	5
3. Mise en place du projet.....	5
3.1. Mise en place des ressources nécessaires.....	5
3.1.1. Ressources humaines.....	5
3.1.2. Moyens informatiques.....	6
3.1.3. Montage et moyens financiers.....	6
3.2. Outils de gestion et de communication du projet et leur usage.....	6
3.3. Coordination des 5 partenaires du projet.....	7
3.4. Infrastructure logicielle.....	8
4. Définition des interfaces entre les systèmes logiciels des partenaires.....	9
5. Implication de chaque partenaire : tâches essentielles réalisées et restant à réaliser	14
5.1. CRDO Paris.....	14
5.1.1. CRDO-Paris : l'application existante.....	14
5.1.2. CRDO-Paris : tâches réalisées dans le cadre du projet pilote.....	14
5.1.3. CRDO-Paris : tâches restant à réaliser.....	15
5.2. CRDO Aix.....	15
5.2.1. CRDO Aix : l'application existante.....	15
5.2.2. CRDO Aix : tâches réalisées dans le cadre du projet pilote.....	15
5.2.3. CRDO Aix : tâches restant à réaliser.....	16
5.3. CINES.....	16
5.3.1. CINES : tâches réalisées.....	16
5.3.2. CINES : tâches restant à réaliser.....	17
5.4. CC IN2P3.....	17
5.4.1. CC-IN2P3 : tâches réalisées.....	17
5.4.2. CC-IN2P3 : tâches restant à réaliser.....	17
6. Réalisation - validation de la chaîne d'archivage principale	18

<u>7. Les transactions supplémentaires.....</u>	<u>18</u>
<u>8. Le développement des outils génériques.....</u>	<u>19</u>
<u>9. Besoins d'archivage des CRN et autres entités.....</u>	<u>19</u>
<u>10. Planning à venir.....</u>	<u>20</u>
<u>11. Interactions avec le contexte global d'Adonis.....</u>	<u>20</u>
<u>11.1. Moteur de recherche global.....</u>	<u>20</u>
<u>11.2. Interactions avec le Méta Portail Adonis.....</u>	<u>20</u>
<u>11.3. Architectures adaptables aux besoins et aux contextes.....</u>	<u>21</u>
<u>11.4. Accès aux données non archivées.....</u>	<u>21</u>
<u>12. Abréviations, liens Internet et références.....</u>	<u>21</u>
<u>12.1. Glossaire.....</u>	<u>21</u>
<u>12.2. Abréviations.....</u>	<u>21</u>
<u>12.3. Liens Internet.....</u>	<u>22</u>
<u>12.4. Références.....</u>	<u>22</u>

1. Objet

Rappelons que le projet pilote s'articule sur trois acteurs essentiels :

- Le CRDO, Centre de Ressources pour la Description de l'Oral, constitué d'un groupe parisien et d'un pôle à Aix-en Provence. Dans la suite de ce document, quand nous parlerons au singulier du CRDO, ce sera pour désigner l'ensemble des deux pôles,
- Le CINES, Centre Informatique National de l'Enseignement Supérieur, à Montpellier,
- Le CC-IN2P3, Centre de calcul de l'Institut national de Physique Nucléaire et de Physique des Particules, à Villeurbanne.

Le TGE ADONIS est le maître d'ouvrage du projet. Un consultant assure la mise en place et le suivi de la maîtrise d'ouvrage.

Ce document d'avancement rend compte de l'organisation mise en place par le projet, des choix essentiels qui ont été retenus, des développements réalisés et des tests effectués. Il dresse également un ensemble de perspectives, tant pour le projet pilote que pour son extension à d'autres entités chargées de l'archivage pérenne d'autres types de données numériques, comme les Centres de ressources numériques.

Il est préférable, pour une bonne compréhension de ce document, de prendre préalablement connaissance du document de synthèse [REF 09] qui définit plus précisément les objectifs du projet, les exigences fonctionnelles générales ainsi que la répartition des responsabilités et des fonctionnalités entre les acteurs.

2. Rappel du schéma fonctionnel du projet pilote

Ce schéma, basé sur le modèle de référence OAIS (Open Archival Information System), a été décrit et analysé en détail dans le document de synthèse. Il est présenté ici pour mémoire.

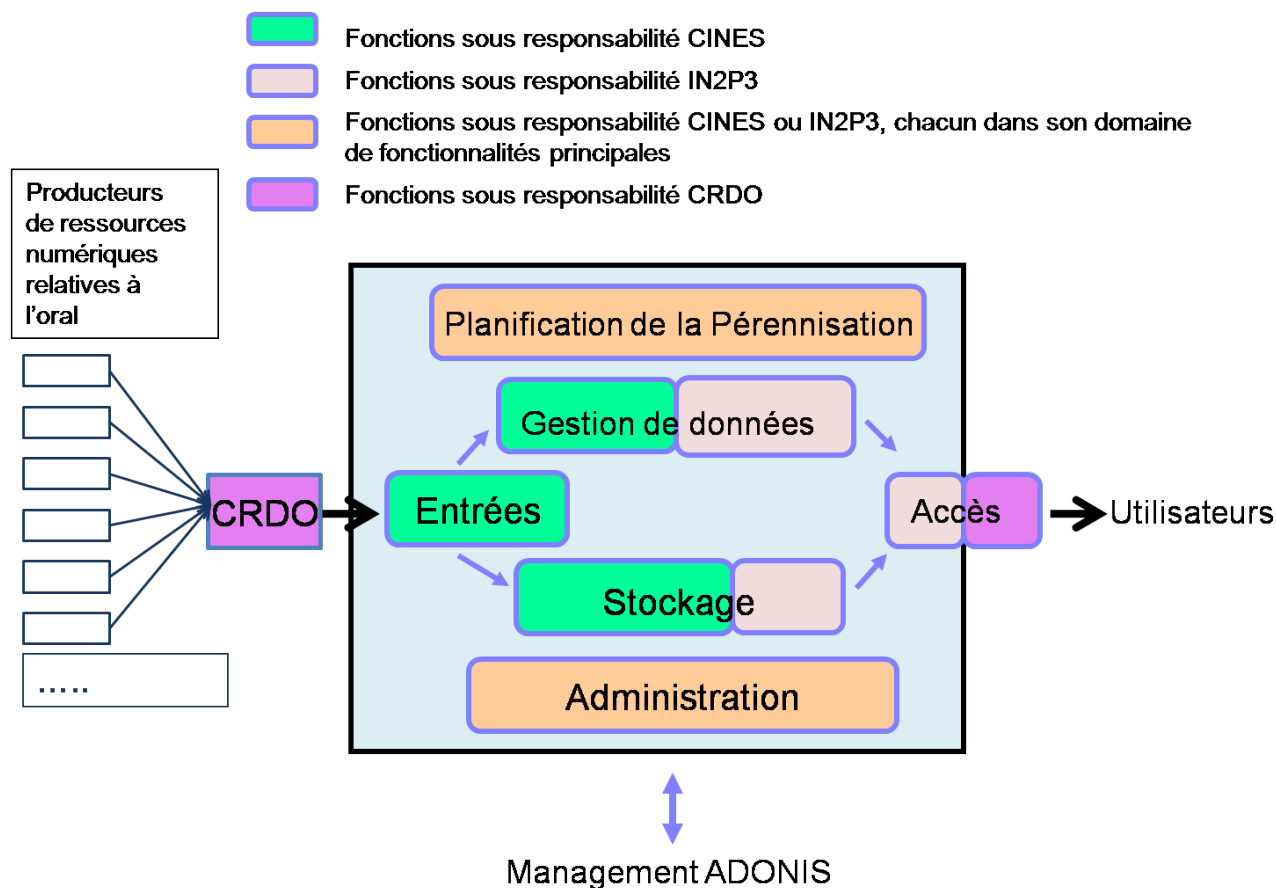


Figure 1: Schéma fonctionnel du projet pilote

3. Mise en place du projet

3.1. Mise en place des ressources nécessaires

3.1.1. Ressources humaines

Courant 2008, la phase initiale de réflexion sur le projet et de clarification du partage des fonctions s'est appuyée sur les ressources internes propres au CINES, au CC-IN2P3 et au CRDO, avec un soutien du TGE Adonis pour les déplacements.

Les travaux plus conséquents de développement de l'infrastructure matérielle et logicielle n'ont pu être effectués qu'à partir de l'embauche sur CDD d'ingénieurs spécifiquement affectés au projet pilote :

- Pierre-Yves Jallud au CC-IN2P3 à partir du 1/9/2008

- Pascal Dugénie au CINES à partir du 1/12/2008

Le CNRS avait créé des Centres de Ressources Numériques (CRN) en 2006, spécialisés par type de document :

- écrit (CNRTL – ATILF, Nancy et CESR, Tours),
- oral (CRDO à Paris et Aix),
- manuscrits (TELMA – IRHT et Ecole Nationale des Chartes),
- images (CN2SV, Paris),
- données géographiques (M2ISA).

Ces CRN n'ont été financés en tant que tels qu'une année. En 2008, certains ont bénéficié d'un soutien indirect du TGE Adonis via l'acceptation de projets du TGE Adonis (appel à projets d'août 2007).

La mission initiale des CRN incluait l'archivage des données. La mise en place du projet pilote d'archivage des données orales et son extension à d'autres données change la nature de cette mission : elle se concentrerait alors sur la relation entre les producteurs de données et le service d'archivage d'une part et entre les utilisateurs de données et le service d'accès aux données archivées d'autre part. Cette mission est à définir précisément et les modalités de sa réalisation et de son financement doivent être déterminées par le TGE Adonis.

3.1.2. Moyens informatiques

Ce sont ceux :

- du RISC (Relais d'information sur les sciences de la cognition) pour la préparation des données à archiver et à rendre accessibles au CRDO Paris,
- du laboratoire Parole et Langage du CNRS pour la préparation des données à archiver et à rendre accessibles au CRDO Aix,
- du CINES pour les fonctions de réception/validation des données à archiver, stockage pérenne, conservation et gestion du fonds, transmission des données au CC-IN2P3,
- du CC-IN2P3 pour tous les aspects relatifs à l'accès aux données : stockage, gestion, protection, recherche, transformation, diffusion.

3.1.3. Montage et moyens financiers

Le principe retenu est celui de la mise à disposition, par le TGE Adonis, de personnel sur les sites du CINES et de l'IN2P3. Cette mise à disposition s'effectue dans le cadre d'une lettre d'intention entre le CC-IN2P3 et le TGE Adonis et dans celui d'une convention en cours de négociation entre le CINES et le TGE Adonis.

Cette solution a pour objectif de permettre un lien étroit entre le maître d'ouvrage du dispositif et les opérateurs chargés de la réalisation. Elle contribue à rendre visibles les coûts effectifs de l'archivage. Elle favorise les évolutions souples et la montée en charge.

3.2. Outils de gestion et de communication du projet et leur usage

Un ensemble d'outils indispensables au bon fonctionnement du projet a été installé par le CC-IN2P3 sur les moyens du CC-IN2P3. Ces sont des outils accessibles à tous ; ils jouent un rôle essentiel dans un projet où les acteurs sont répartis sur 6 sites géographiques distincts (TGE-Adonis, CINES, CC-IN2P3, CRDO Paris, CRDO Aix, RISC) :

- la liste Adonis-Archivage est une liste de diffusion de courrier électronique incluant l'ensemble des personnes impliquées dans le projet : adonis-archivage-l@in2p3.fr,
- le Wiki du projet pilote (<http://www.tge-adonis.fr/wiki/index.php/Accueil>). Il est basé sur le logiciel Open source MediaWiki et joue deux rôles :

- c'est un espace de travail collaboratif permettant aux acteurs du projet de coordonner leurs activités, de disposer en permanence d'informations de référence à jour, etc.,
- c'est un espace qui est public pour l'essentiel et qui offre aux CRN (Centres de ressources numériques) mais aussi à tous les acteurs concernés par la mise en place de solutions d'archivage mutualisées, une visibilité sur l'avancement du projet pilote.
- un outil de gestion de projet (Dotproject) permettant de gérer la définition et l'ordonnancement des tâches. Tous les partenaires du projet n'ayant pas la même culture ni la même expérience en matière de gestion de projet, l'usage de cet outil reste insuffisant,
- la plate-forme IN2P3 de téléconférence est également un outil d'usage très régulier.



Figure 2 : Page d'accueil du Wiki du projet pilote

3.3. Coordination des 5 partenaires du projet

Les cinq partenaires sont le TGE Adonis, le CRDO Paris, le CRDO Aix, le CINES et le CC IN2P3.

La coordination passe par trois canaux principaux :

- les réunions périodiques de l'ensemble des partenaires. La périodicité moyenne est de l'ordre de 6 semaines (7 réunions d'une journée ont été tenues),
- les téléconférences audio de niveau projet toutes les deux semaines, parfois plus souvent (12 téléconférences d'une heure trente ont été tenues),
- les outils de communications communs décrits ci-avant.

Réunions et téléconférences donnent lieu à des comptes rendus systématiques en ligne sur le Wiki.

3.4. *Infrastructure logicielle*

L'infrastructure mutualisée de pérennisation et d'accès aux données s'appuie sur plusieurs systèmes logiciels essentiels :

La PAC (Plate-forme d'archivage du CINES) est une plate-forme d'archivage numérique opérationnelle qui a été mise en place notamment pour l'archivage des thèses des universités françaises (cf. 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 du doctorat). Cette plate-forme s'appuie elle-même sur l'application logicielle d'archivage ARCSYS d'Infotel.

Dans le cadre du projet pilote, cette plate-forme doit faire l'objet d'un ensemble d'évolutions permettant de prendre en compte le contexte particulier de l'archivage de données scientifiques (élargissement de la nature des données, méthodologies de recherche, relations entre objets, versionnage, droits d'accès, protection des données personnelles...).

Pour satisfaire ces besoins spécifiques, le CINES a entrepris de faire évoluer sa plateforme d'archivage PAC avec les adaptations suivantes:

- gestion d'une relation de parenté entre documents archivés (les enregistrements oraux donnent lieu, sont conceptuellement « père » ou « mère », des annotations – transcriptions – qui portent sur eux ; les données et leurs annotations sont regroupés en « collections » qui sont conceptuellement leur « père » ou leur « mère »),
- allocation dans ce qui est déposé d'un espace de métadonnées descriptives propres à chaque métier,
- mise en place d'un mécanisme de mise à jour des métadonnées descriptives sans re-transfert complet du document archivé (pour gérer par exemple l'évolution des droits d'accès au document),
- relais vers la partie Accès d'informations de diffusion présentes dans ce qui est déposé et ignorées par le processus d'archivage,
- intégration de la validation de nouveaux formats audio-vidéo éligibles à l'archivage.

Le produit iRods (i Rule Oriented Data Systems, <https://www.irods.org/>) est un outil Open source qui fournit une palette riche d'outils d'aide à la gestion de données. Il permet de virtualiser la gestion de données et l'accès à des ressources de stockage réparties. La virtualisation vise à rendre les fonctions de base de stockage, de transfert de données, de contrôle d'intégrité, etc., indépendantes des technologies, supports physiques d'enregistrements, protocoles... utilisés aujourd'hui ou demain pour réaliser ces opérations. De plus, ce produit intègre un gestionnaire de règles pouvant s'appliquer à tout ou partie des composantes du système, comme les données, les utilisateurs, les ressources...

Le produit iRods est utilisé pour opérer les transferts de données entre le CINES et le CC-IN2P3. Il pourra également être utilisé entre les CRN et le CINES pour les transferts massifs de données.

Il facilitera également les opérations automatiques sur les données suite aux transferts en provenance du CINES (engendrement de versions de diffusion, etc.).

Le logiciel Fedora Commons (<http://www.fedora-commons.org/>) est un Open source qui couvre une large palette de fonctionnalités d'archivage numérique parmi lesquelles, on peut trouver des mécanismes d'organisation des données, la gestion de la description de ces données par des métadonnées, la recherche d'information, etc. Ce logiciel est utilisé à l'IN2P3 dans le cadre du projet pilote afin de couvrir les fonctionnalités liées à l'organisation des données en vue de l'accès. Fedora Commons est utilisé aujourd'hui par des bibliothèques, des universités, des institutions relevant du domaine culturel, dans des projets similaires au plan international (comme le projet ESFRI DARIAH), etc. Bien que la BnF ait récemment renoncé à l'usage de Fedora pour des raisons de performances (plusieurs pétaoctets de données, des millions de référence archivées), il a cependant été considéré que ce produit pouvait tout à fait répondre aux besoins du projet pilote et de son élargissement futur aux autres données des SHS. Ce produit utilise, au CC-IN2P3, iRods comme système de gestion de données et d'accès à ces données au niveau fichiers. Fedora-Commons

permet, quand à lui, une gestion au niveau d'objets numériques pouvant être composés de plusieurs fichiers de données et de métadonnées.

4. Définition des interfaces entre les systèmes logiciels des partenaires

Dans un projet de cette nature, la question de la définition des interfaces, c'est à dire de la définition des contenus, des formats, des protocoles, des procédures entre les systèmes logiciels des partenaires est essentielle pour que la chaîne de pérennisation et d'accès aux données fonctionne.

<i>Interface entre</i>	<i>remarques</i>
producteur de données et CRDO	<p>L'existence d'une entité intermédiaire de pré-versement entre les chercheurs, les projets, les équipes producteurs de contenus scientifiques d'une part et l'infrastructure technique de pérennisation et d'accès constituée par le CINES et le CC-IN2P3 est fondamentale.</p> <p>Le rôle du CRDO dans ce cadre est double :</p> <ul style="list-style-type: none"> • il assure la collecte des objets numériques auprès des producteurs, il procède à une mise en forme de ces objets en paquets archivables et conformes aux exigences de la conservation à long terme, il dispose d'une connaissance scientifique suffisante pour assurer un certain contrôle de qualité sur les contenus et il assure le versement des paquets au CINES. • il conduit également un travail en direction des producteurs de contenus scientifiques : sensibilisation aux risques de perte du patrimoine numérique, incitation à l'archivage et à la mise à disposition rapide des données, diffusion des recommandations essentielles pour l'archivage (formats de données recommandés), diffusion des outils logiciels, des standards et des bonnes pratiques dans ce domaine. <p>Les interfaces entre les producteurs de données et le CRDO ne donnent pour l'instant pas lieu à un formalisme particulier. Le CRDO dispose d'une compétence technique sur les formats de collecte ainsi que sur ceux de diffusion (transformation, contrôle, etc.) et sur les outils associés pour leur mise à disposition.</p> <p>Dans la pratique, nous ne disposons que de peu de visibilité et de peu de recul sur ces rôles du CRDO car dans la première phase de travail, on archivera d'abord les objets numériques qui ont été collectés depuis un certain temps par le CRDO.</p>
CRDO et CINES	<p>C'est l'entrée des données à archiver dans l'infrastructure de pérennisation et d'accès. Les contraintes techniques et scientifiques doivent ici être définies dans le détail.</p> <p>Sur le plan technique, les données à verser sont organisées sous forme de paquets d'informations à verser (les SIP, Submission Information Package) constituée d'un fichier de métadonnées de préservation (fichier sip.xml), d'un répertoire contenant les données et métadonnées à conserver (répertoire "depot") et d'un répertoire réservé aux informations de diffusion qui n'ont pas lieu d'être pérennisées. La figure 3 présente un schéma du paquet d'informations à verser.</p> <p>Le processus de transfert du SIP entre le CRDO et le CINES, ainsi que les opérations de vérification du SIP par le CINES et les différents messages transmis par le CINES au CRDO dans ce cadre (accusé de réception, message d'anomalie, certificat d'archivage confirmant la validation du SIP et sa prise en compte par le CINES) sont spécifiés en détail sur le Wiki : http://www.tge-adonis.fr/wiki/index.php/Interface_de_versement_au_CINES</p>

	<p>La figure 4 présente un synoptique global de ce processus d'archivage.</p> <p>Sur le plan scientifique, la signification précise, dans le cadre du projet pilote, d'un certain nombre d'éléments de contenu et le mode de représentation de ces contenus ont été définis. Le vocabulaire contrôlé sera élargi progressivement avec les autres CRN.</p> <p>Se reporter à http://www.tge-adonis.fr/wiki/index.php/Actions_et_r%C3%A9flexions_communes_aux_CRDO</p>
CINES et CC-IN2P3 -	<p>Cette interface est essentiellement basée sur la mise en œuvre de iRods. Un ensemble de serveurs iRods est installé au CC-IN2P3 dans le cadre du projet pilote. Un client iRods est installé au CINES. Une application développée par le CINES permet de lancer automatiquement le transfert du paquet archivé (l'AIP) vers le CC-IN2P3.</p> <p>L'AIP a une structure similaire à celle du SIP. Lors du processus d'archivage, le SIP a été complété par un certain nombre d'informations supplémentaires (affectation d'identifiant, calcul d'empreintes numériques...)</p> <p>Le détail de l'interface est spécifié sur http://www.tge-adonis.fr/wiki/index.php/Interface_CINES-IN2P3</p>
CC-IN2P3 et application métier	<p>Le CC-IN2P3 ayant vérifié l'intégrité les données en provenance du CINES et les ayant ingéré dans Fedora-Commons, les objets numériques ainsi formés, composés de données et de métadonnées, sont accessibles par les applications métiers des CRDO au travers de web services basés sur le protocole SOAP. Ces web services satisfont le protocole OAI-PMH permettant les recherches de nouveaux objets et le moissonnage. La figure 5 présente un aperçu des fonctionnalités d'accès.</p>
application métier et utilisateur final	<p>Il s'agit des applications web développées par le CRDO Paris ou le CRDO Aix qui offrent aux utilisateurs finaux une interface web pour rechercher des données et les récupérer. A noter que d'autres interfaces normalisées existent pour l'accès aux données, comme par exemple l'utilisation du protocole OAI-PMH.</p> <p>Dans un premier temps, les utilisateurs ne devraient pas voir la différence par rapport à l'application pré-existante, bien qu'ici, les données ne soient plus conservées en local mais transmises à l'application par l'infrastructure du CC-IN2P3.</p>

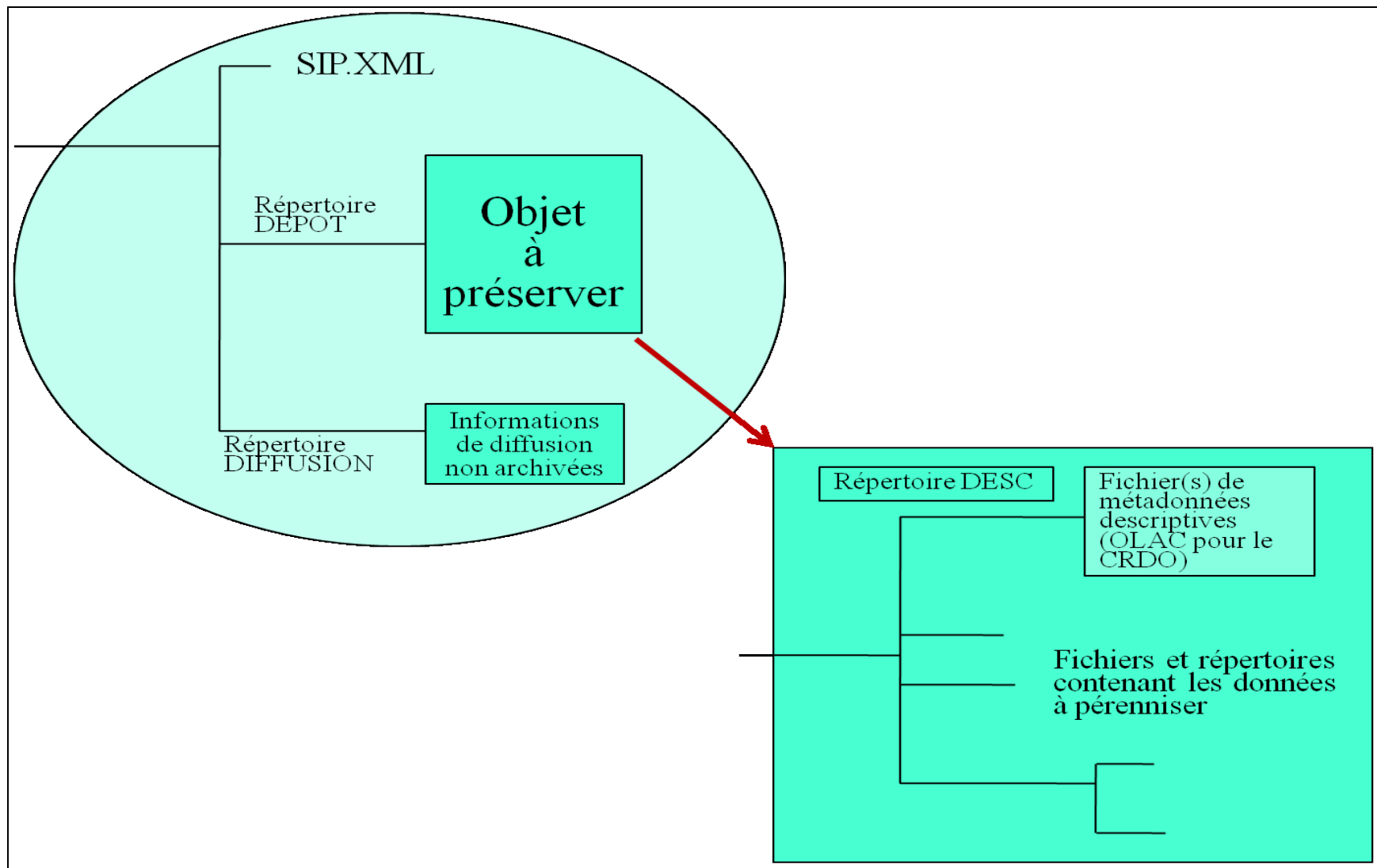


Figure 3 : Schéma du paquet d'informations à verser (SIP)

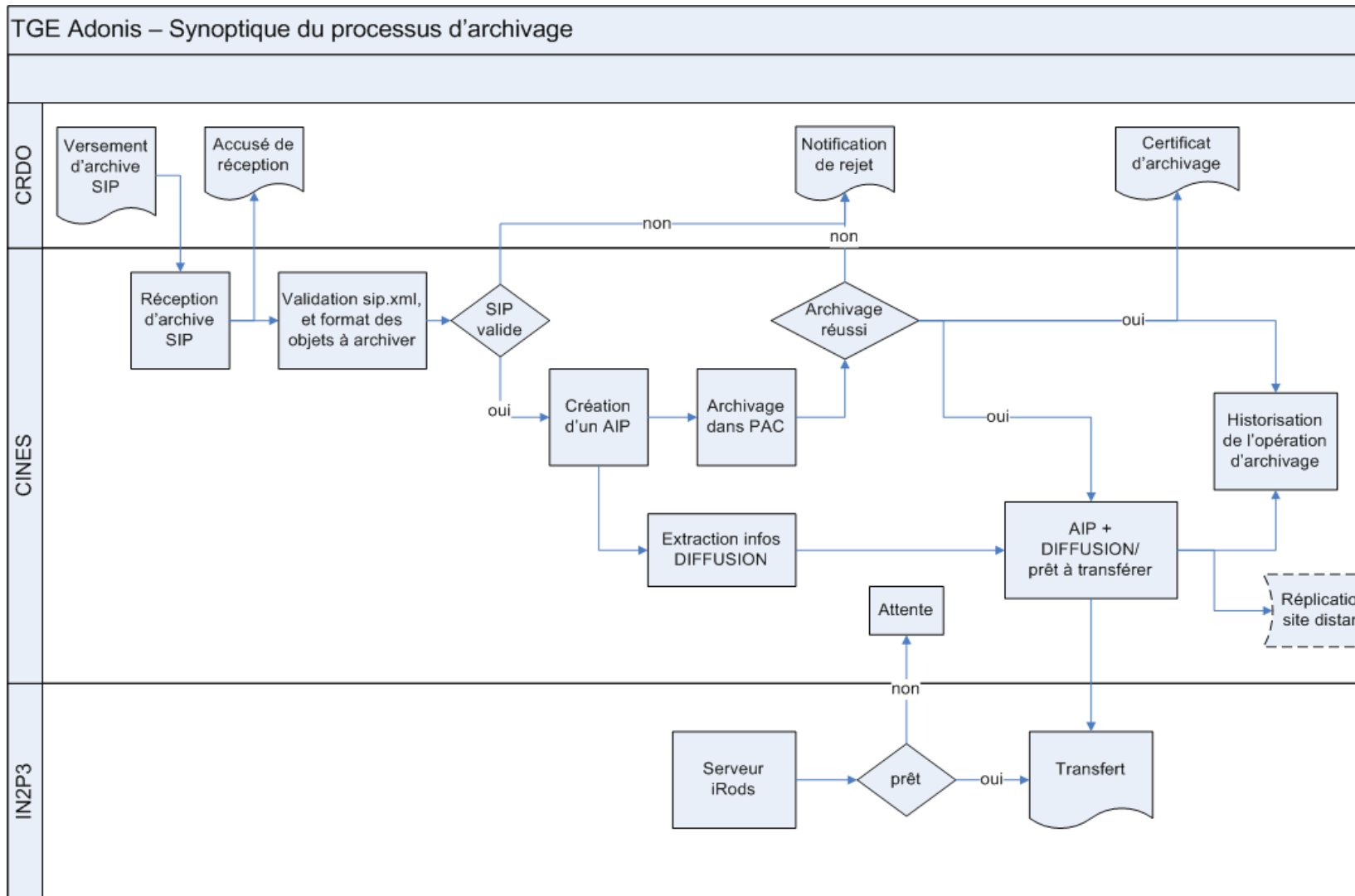


Figure 4 : synoptique du processus d'archivage

5. Implication de chaque partenaire : tâches essentielles réalisées et restant à réaliser

5.1. CRDO Paris

5.1.1. CRDO-Paris : l'application existante.

L'application du CRDO Paris (<http://crdo.risc.cnrs.fr/>) est actuellement hébergée par le RISC.

- **Accès aux ressources par du moissonnage OAI.**

Les fournisseurs de services (OAIster, Scientific Commons), moissonnent régulièrement le réservoir du CRDO et offrent des services de type moteur de recherche dans des ensembles plus vastes de ressources.

Des portails institutionnels, communautaires, de projet ou d'équipe moissonnent tout ou partie du réservoir pour présenter les ressources de différentes manières pour différents publics. Par exemple le site corpusdelaparole.culture.fr présente une collection de ressources sur le français et les langues de France, ce portail devrait très prochainement être moissonné à son tour pour alimenter le guichet unique Culture.fr. Autre exemple, le portail OLAC moissonne plus d'une trentaine d'archives sur les langues, dont le CRDO et constitue un catalogue accessible aussi par le site de LinguistList.

- **Accès par le portail CRDO (<http://crdo.risc.cnrs.fr/>).**

Ce portail permet d'accéder à l'ensemble des ressources publiques comme à celui des ressources en cours de construction. Une gestion des droits d'accès est mise en place pour ce dernier ensemble.

Des moteurs de recherche exploitant les métadonnées sont proposés (par mot-clé, par catégories Dublin-Core, par une carte géographique et par un axe temporel)

Les ressources sont toutes téléchargeables. Les données vidéo sont aussi consultables en streaming.

Différentes vues sont proposées pour consulter simultanément un enregistrement et sa transcription.

Des outils de recherche dans les transcriptions sont proposés (recherche d'occurrences de mots)

Un outil de consultation d'extraits sonores permet de construire ces extraits à la volée

5.1.2. CRDO-Paris : tâches réalisées dans le cadre du projet pilote

Sélection des ressources : il a été décidé de ne travailler dans un premier temps que sur des ressources simples composées d'un enregistrement audio et accompagnées éventuellement d'un fichier d'annotations (transcriptions, traductions, etc.) dans un format XML ou PDF. Nous nous limiterons aussi dans ce premier temps aux ressources déclarées par les déposants comme librement accessibles.

Développement d'un outil de vérification de la bonne formation des formats fichiers WAV.

- vérification des contraintes de codage, de formatage ainsi que des contraintes de qualité (fréquences d'échantillonnage, taille de l'échantillon, présence de plages silencieuses),
- suppression de toutes les informations autres que celles sur lesquelles on engage la responsabilité de la conservation (Il s'agit en général d'informations non formalisées et non normalisées qui ont momentanément été utiles au producteur).

Conversion des anciennes DTD utilisés en schémas XML et affectation d'un espace de noms.

Développement d'un script de construction de paquet de versement. Ce script effectue les premières vérifications, prépare les paquets au format défini et verse le résultat au CINES. Validation, en interaction avec le CINES, du processus de versement par ce script.

Adaptation d'un programme de création à la volée d'extraits sonores afin de lui permettre d'accéder à des fichiers distants à travers l'interface iRods du CC-IN2P3.

5.1.3. CRDO-Paris : tâches restant à réaliser

Processus de mise à jour de la base de données documentaire afin de modifier les caractéristiques des ressources archivées (déclaration du statut archivé, déclaration du numéro de version, affectation de l'identifiant ARK et changement de l'URL d'accès) pour un accès non plus en local mais via le gestionnaire FEDORA-commons distant.

Ces choix seront faits en ayant en perspective la migration complète de l'application « métier » en utilisant les ressources de l'IN2P3. Dans le même ordre d'idée, il faut dès à présent réfléchir aux outils de conversion de format (de celui d'archivage vers celui de diffusion) de la manière la plus automatisée possible ainsi que la mise en place sur la plate-forme de l'IN2P3 d'outils de diffusion adaptés (e.g. streaming).

Processus de d'automatisation et de gestion de l'enchaînement des versements (traitement des messages en retour du CINES, définition des relations entre objets...) Il s'agit ici de verser plusieurs milliers d'objets.

5.2. CRDO Aix

5.2.1. CRDO Aix : l'application existante

Les caractéristiques essentielles de l'application existante (<http://crdo.fr/>) doivent être reconduites dans le contexte de l'archive mutualisée :

- **Accès intégralement multilingue.** Actuellement, 4 langues : chinois, espagnol, anglais, français.

La *diffusion* des corpus (streaming audio ou vidéo) est possible (en intégralité ou partiellement) sans identification préalable sur le site. Par contre, la *distribution* des objets (corpus et ressources qui leur sont associées) est conditionnée par la signature préalable d'une **licence utilisateur** (cf. http://crdo.fr/wiki/Licences_fr) et la restriction (définie par le producteur) à certaines **catégories d'utilisateurs**. Dans la plupart des cas, la distribution de corpus étiquetés et d'enrichissements est réservée à la communauté de l'enseignement et de la recherche. Cette restriction est destinée à empêcher toute utilisation des ressources à des fins autres (commerciales par exemple) que la recherche dans un contexte de communication scientifique directe.

- **Traçabilité des opérations de téléchargement et suivi des utilisateurs.** Ce suivi permet à tout utilisateur identifié sur le site de visualiser les téléchargements d'un objet et d'entrer en contact avec les bénéficiaires de ces téléchargements. Ce dispositif encourage l'émergence de communautés de producteurs et d'utilisateurs susceptibles de collaborer sur des programmes de recherche faisant le meilleur usage des ressources disponibles. Par ailleurs il permet de repérer plus facilement toute utilisation non conforme aux engagements de la licence utilisateur.
- **Affichage des publications** associées à un objet particulier, ainsi que de ses **relations** avec des équipes, des laboratoires, des programmes de recherche et/ou des réseaux thématiques. Ces affichages permettent à la fois de mieux caractériser l'objet et de rendre compte de son usage dans diverses activités de recherche.
- **Espaces de travail** pour les projets associés à certains corpus, avec accès privilégié aux données et possibilité de mettre à jour de manière incrémentale les enrichissements distribués. Depuis février 2009 l'accès à ces espaces de travail n'est plus restreint à l'intranet du LPL.

5.2.2. CRDO Aix : tâches réalisées dans le cadre du projet pilote

« Mise à plat » de l'arborescence des répertoires et renommage des fichiers pour récupération directe dans l'environnement FEDORA. Correction automatique des liens sur les fichiers HTML en fonction de cette nouvelle structuration.

Conversion automatique de fichiers au format AIFF (Audio Interchange File Format, format de fichier audionumérique développé par Apple) vers le format WAV (fichier audionumérique élu pour l'archivage).

Génération automatique du fichier de description du paquet SIP et du fichier de métadonnées descriptives. Les relations de parenté entre objets sont formées à partir des relations présentes dans le schéma Dublin Core de chaque objet.

Versement automatique d'une file d'objets par lancement d'un script.

Récupération automatique de l'identificateur unique (ARK, Archival Resource Key) par analyse du descripteur de l'AIP contenu dans le message de confirmation de l'archivage (certificat d'archivage). Cet identificateur est utilisé pour remplacer automatiquement les adresses locales des objets à télécharger par leurs adresses dans l'environnement FEDORA du CC-IN2P3.

5.2.3. CRDO Aix : tâches restant à réaliser

Conversion automatique de fichiers graphiques vectoriels au format EPS (non accepté pour l'archivage) vers le format SVG (élu pour l'archivage).

Conversion de fichiers au format texte de l'application PRAAT (application logicielle open source d'analyse de fichiers son) vers le format XML : ce développement sera confié au développeur de PRAAT à partir de spécifications fournies par le CRDO Paris.

Traitement de données particulières : enregistrements sonores multicanaux, données articulatoires (station EVA), données électrophysiologiques (Potentiel évoqué), images IRMf (imagerie par résonance magnétique fonctionnelle), etc. à reformater pour permettre leur archivage pérenne chaque fois que cela est souhaitable.

Spécification des formats souhaités pour le stockage et l'archivage de documents vidéo, avec procédures de transcodage si nécessaire.

Evolutions de l'application d'accès pour pouvoir accéder aux fichiers archivés. L'application d'accès restera localisée à Aix dans un premier temps.

5.3. CINES

5.3.1. CINES : tâches réalisées

Protocole de versement : Le nouveau modèle conceptuel du SIP est validé, ainsi que le protocole de versement (tel que décrit par le diagramme de flux) en provenance du CRDO Paris et du CRDO Aix. Ce paquet et les différentes transactions associées constituent une extension par rapport à la chaîne d'archivage existant au CINES.

Déploiement sur PAC : Après une série de tests du processus complet réception/validation/constitution du paquet archivé/transfert au CC-IN2P3 sur un ensemble significatif de cas standards pour des enregistrements audio et des annotations correspondantes, l'outil d'analyse et de validation du SIP pour le projet TGE Adonis a été déployé sur la plateforme PAC. Une plateforme de test autonome avec des schémas sip.xsd et aip.xsd séparés a été mise en place pour éviter tout risque de perturbation de la plateforme de production.

Nouveaux formats : Une étude analysant les choix des formats audio et vidéo éligibles pour l'archivage a été menée. Dans sa nouvelle version, PAC intègre le module de validation des formats audio "Wav" (avec encodage PCM).

Transfert des paquets archivés au CC-IN2P3 : Le transfert des paquets archivés au CINES vers le CC-IN2P3 est opérationnel au travers d'un client iRods.

Diffusion : Les informations de diffusion fournies par les CRDO sont convenablement relayées à l'IN2P3 dans un répertoire ignoré par le processus d'archivage.

Principal problème rencontré : L'utilisation de iRods a demandé un effort important pour maîtriser l'usage des bibliothèques et tester les différentes solutions possibles. Actuellement nous ne pouvons pas nous satisfaire d'une solution qui rendrait le processus de versement totalement synchrone avec le processus de transfert entre le CINES et le CC-IN2P3. Or, un processus asynchrone complexifie notablement le processus dans son ensemble et remet en question certains choix initiaux (ex. l'usage de la bibliothèque java jargon).

5.3.2. CINES : tâches restant à réaliser

Protocole de versement : Il s'agit principalement de finaliser le processus de mise à jour des métadonnées descriptives. Également, il est souhaitable d'améliorer la précision des avis de rejet renvoyés aux services versants et d'intégrer de nouveaux modules de validation de formats éligibles à l'archivage.

Sauvegarde distante : Une sauvegarde des archives est actuellement effectuée sur le site du CINES. Toutefois, afin de minimiser les risques de perte de données en cas de sinistre local, le CINES entrevoit de réaliser une sauvegarde distante à la BnF. Cet aspect est en cours de discussion avec la BnF, un calendrier prévisionnel pour sa mise en place au début de 2010 est à l'étude. En attendant, une copie de la sauvegarde effectuée est transmise de façon hebdomadaire à un organisme voisin - le LIRMM - dont les bâtiments sont distants d'environ 500m du site du CINES.

En outre, le CC-IN2P3 effectue une sauvegarde des paquets archivés qu'il reçoit du CINES.

Interface CINES-IN2P3 : Renforcement du contrôle de la synchronisation des archives entre le CINES et le CC-IN2P3 en cas d'interruption de service ou de rupture de connexion (traitement des erreurs, gestion des temporisations hors délai).

5.4. CC IN2P3

5.4.1. CC-IN2P3 : tâches réalisées

Un important travail de compréhension des besoins et des spécificités de la communauté SHS a été réalisé par le CC-IN2P3 afin de pouvoir entreprendre de manière optimale les déploiements matériels et logiciels nécessaires.

Mise en place et mise en configuration des machines dédiées à l'application Fedora-Commons (gestion des données – dans leurs différentes « incarnations » numériques – et de leurs métadonnées) et à l'application LDAP (authentification des utilisateurs en fonction d'appartenances et de profils) permettant d'avoir des environnements de test, de développement et de production.

Mise en œuvre d'une instance d'iRods spécifique à Adonis basée sur :

- un serveur de données de 40 TO.
- un serveur gérant le méta-catalogue de cette instance et connecté à un cluster Oracle spécifique pour la partie base de données.

Définition du modèle de données permettant de mettre en correspondance les constituants du paquet archivé transmis par le CINES et les objets Fedora.

Développement des scripts et règles permettant la vérification de l'intégrité des données et la construction des objets Fedora à partir de l'AIP (le « paquet » archivé) envoyé par le CINES.

5.4.2. CC-IN2P3 : tâches restant à réaliser

Intégration de services à plus-value ajoutées dans le système comme la diffusion de données dans d'autres formats que ceux de l'archive, l'accès aux données audio et vidéo par streaming, ... Cette liste de services est encore à définir avec les différents acteurs du projet.

Intégration de l'ensemble des transferts de fichiers et pilotage de ces transferts (CRDO-CINES) à partir d'un espace de travail collaboratif construit sur iRods.

Construction d'un environnement global permettant l'authentification, la gestion des droits et des identités, cohérent entre l'ensemble des différentes applications constituantes du projet et de ses possibles extensions. Cette construction permettra également de déployer un environnement plus sécurisé, basé sur des protocoles d'encryptage des données, des droits et des identités lors des transferts et communications entre applicatifs. Un catalogue moissonnable de l'ensemble des métadonnées SHS sera également constitué.

6. Réalisation - validation de la chaîne d'archivage principale

Cette chaîne part de la constitution de paquets de données à archiver au CRDO et se termine par la vérification de l'accès à ces mêmes données après archivage depuis l'application d'accès. Elle comporte les étapes successives suivantes :

1. Création des paquets de versement par le CRDO (développement d'un script de création automatique des paquets à partir des données existantes)
2. Versement des paquets au CINES
3. Réception/validation de ces paquets, affectation d'un identifiant, constitution de l'AIP
4. Transmission du certificat d'archivage au CRDO
5. Mise en place du processus de conservation de l'AIP
6. Transmission de l'AIP à l'IN2P3
7. Organisation des différents constituants du paquet au sein d'une base de données Fedora, récupération des métadonnées
8. *Transmission de toutes les informations utiles pour l'application métier ouverte aux utilisateurs finaux*
9. *Validation du fonctionnement de cette application*

Fin mars 2009, cette chaîne est réalisée jusqu'à l'étape 7 (les étapes restant à réaliser sont en italiques).

Restent à mettre en place la communication entre l'infrastructure générale d'accès aux données du CC-IN2P3 et les applications métier spécifiques du CRDO.

7. Les transactions supplémentaires

Il s'agit de transactions CRDO-CINES (puis par voie de conséquence CINES-IN2P3 puis IN2P3-application CRDO) qui visent à compléter la chaîne d'archivage principale.

Ces transactions correspondent à des besoins exprimés pour les données issues de la recherche. Elles sont générales et non spécifiques au CRDO. Ce sont les suivantes :

- mise à jour de métadonnées descriptives d'un objet déjà archivé (correction, amélioration...),
- changement des contraintes et restrictions d'accès pour un objet ou une collection d'objets,
- gestion des versions d'un même objet. En pratique, ce point n'implique pas de transaction de versement spécifique mais il implique de renseigner de façon adéquate les métadonnées permettant cette gestion.
- gestion de collections d'ensembles d'objets. Là encore, il n'y a pas de transaction de versement spécifique. Un premier versement permet d'archiver un objet "collection". Il est alors possible de verser des objets et de préciser dans les métadonnées que ces objets appartiennent à telle ou telle collection. L'opération consistant à créer une collection correspondant à un ensemble d'objets déjà archivés est également possible mais plus lourde que la précédente.

La plate-forme du CINES ne permettait pas de traiter les deux premiers cas. Les interfaces de versement ont été enrichies afin de prendre en compte ces besoins.

Ces transactions doivent également être traitées de façon spécifique par le CC-IN2P3 et les applications CRDO.

8. Le développement des outils génériques

Nous avons vu que la structure du SIP et l'identification des différentes transactions de versement avaient été définies de manière à répondre au mieux aux besoins généraux d'archivage des données de la recherche et pas seulement ceux du CRDO.

La démarche est similaire du côté des outils de recherche et d'accès aux données : l'objectif est de réduire les applications métier (du CRDO ou de tout autre CRN) à ce qui est strictement métier et donc d'éviter toute duplication de fonctions entre ces différentes applications.

Ce qui est fait aujourd'hui sur cet aspect :

- le stockage des données à diffuser est assuré dans la base Fedora-Commons avec un niveau d'abstraction au niveau de l'objet numérique et non plus au niveau des applications métier.

Ce qui est prévu à court terme pour le projet pilote :

- la gestion et la propagation des authentifications liées aux restrictions d'accès, sont gérées par un annuaire LDAP commun,
- la mise en œuvre de logiciels de transformation entre format d'archivage et formats de diffusion.

Perspectives à approfondir :

- explorer les possibilités de l'outil Fedora sur ces aspects.

9. Besoins d'archivage des CRN et autres entités

La réunion du 9 octobre 2008, tenue avec l'ensemble des Centres de ressources numériques (compte-rendu accessible sur <http://www.tge-adonis.fr/wiki/uploads/1/15/CRReunion-20081009.doc>) a permis d'établir une première évaluation des besoins. Ces besoins existent, avec des volumétries et des degrés d'urgence variables pour tous les CRN ainsi que pour ArcheoVision (modélisation 3D).

ArcheoVision et CN2SV s'appuient déjà sur les moyens de l'IN2P3, soit pour assurer le stockage de volumes importants de données, soit pour l'hébergement de serveurs web d'accès aux données. Ces deux entités pourraient probablement entrer rapidement dans ce dispositif d'archivage. Les besoins du CNTRL dans ce domaine sont également très importants.

Par ailleurs, certaines entités ont des besoins majeurs d'espace de stockage :

- des projets fortement numériques qui fabriquent des données et qui ont besoin de les stocker, de les charger/décharger aisément, de les partager, etc. C'est typiquement le cas des projets ANR,
- des laboratoires ont des fonds numériques dont ils veulent avoir un "miroir" pour être sûrs de ne pas les perdre, dans la mesure où, même s'ils ont une solution de type RAID, ils n'ont qu'un exemplaire de leurs données.

Il y a lieu, dans le cadre d'Adonis, d'apporter une réponse à ces besoins mais il convient en parallèle de considérer que ces services de stockage correspondent à une situation transitoire visant aussi à préparer l'archivage pérenne.

Plus généralement, un travail de sensibilisation en direction de la communauté SHS doit être entrepris sur ce terrain : sensibiliser les chercheurs par rapport aux risques encourus de perte partielle ou totale de leur patrimoine informationnel, valorisation de leur travaux et protection des droits de propriété intellectuelle correspondants.

10. Planning à venir

	Poursuite de la mise en œuvre de l'archivage dans le cadre du projet pilote	Élargissement du projet pilote
Mai 2009	Finalisation et validation de l'infrastructure matérielle et logicielle : - transactions spécifiques (mise à jour des métadonnées) - prise en compte des données vidéo, - traitement des versions, des collections - gestion centralisée des authentifications, règles d'accès, licences... - adaptation des applications métier du CRDO	Identification des critères de choix des CRN et autres entités candidates à l'archivage ¹
Juin 2009		
Juillet 2009		
Aout 2009		
Septembre 2009	Validation de la chaîne opérationnelle	Lancement du travail préparatoire avec la ou les entités prioritaires
Octobre 2009	Lancement de l'archivage de masse définitif pour les deux CRDO	Planification et organisation du travail en vue de l'archivage
Novembre 2009		
Décembre 2009	Rodage du fonctionnement de routine	

11. Interactions avec le contexte global d'Adonis

11.1. Moteur de recherche global

Les métadonnées utilisées par les CRN relèvent de normes ou de standards propres aux communautés concernées : métadonnées OLAC pour le CRDO, métadonnées ISO 19115 pour les données géographiques, TEI pour le CNRTL, EXIF (Exchangeable image file format) et IPTC (norme IPTC photo metadata 2008) pour le CN2SV...

Une première réunion avec les CRN avait permis d'identifier le Dublin Core comme schéma minimal de métadonnées commun à tous. La disponibilité de ces métadonnées dans le paquet de versement ouvre la possibilité d'activer sous Fedora, un moteur de recherche sur l'ensemble des données SHS archivées.

Il est par ailleurs envisagé d'aller plus loin dans ce domaine en exploitant en profondeur, les schémas de métadonnées propres à chacun des domaines des SHS : chaque fois qu'un fichier de métadonnées sera rattaché à un schéma XML, le moteur de recherche de Fedora pourra exploiter le contenu de ces fichiers.

11.2. Interactions avec le Méta Portail Adonis

Ce point reste à approfondir mais on peut d'ores et déjà prévoir :

- que les moteurs de recherche d'information sur l'ensemble des données SHS disponibles seront accessibles depuis le Méta Portail Adonis,
- que les services d'espaces collaboratifs du méta Portail pourront être utilisés dans le processus d'archivage, ce qui permettra d'éviter la dispersion de ces espaces dans tous les laboratoires et la

¹ On peut en outre penser à l'intégration des « gros » producteurs de données orales : projet PFC et laboratoires Modyco et CLLE-ERSS, projet CLAPI à ICAR.

multiplication des efforts d'administration et de maintien en fonctionnement.

11.3. Architectures adaptables aux besoins et aux contextes

Dans le contexte du projet pilote, deux architectures sont mises en œuvre sur la base de la même infrastructure :

L'application du CRDO Paris a été portée sur les moyens informatiques du CC-IN2P3. Le CRDO Paris y trouve des avantages en termes de maintien en fonctionnement opérationnel des machines et des systèmes d'exploitation.

L'application du CRDO Aix va fonctionner, tout au moins dans un premier temps, sur son site actuel. Cette application communiquera avec l'infrastructure d'accès du CC-IN2P3 avec les mêmes protocoles de communication que dans le fonctionnement qui prévalait avant la mise en place du projet d'archivage.

On peut observer ici une capacité essentielle du dispositif à s'adapter aux conditions spécifiques de telle ou telle communauté.

11.4. Accès aux données non archivées

Dans leur nouveau contexte, les applications du CRDO permettent d'accéder aux données archivées mais aussi à des données qui ne le sont pas car elles n'ont pas encore atteint le degré de stabilité requis pour leur archivage.

Cet accès unifié à toutes les données place l'archivage pérenne au sein du processus vivant de création et d'utilisation des données et des documents scientifiques.

12. Abréviations, liens Internet et références

12.1. Glossaire

On se reportera, si besoin, au glossaire en ligne sur le Wiki du projet pilote : <http://www.tge-adonis.fr/wiki/index.php/Glossaire>

12.2. Abréviations

Abréviation	Nom détaillé
AIFF	Audio Interchange File Format
BnF	Bibliothèque nationale de France
CINES	Centre informatique national de l'enseignement supérieur
CNRS	Centre national de la recherche scientifique
CNRTL	Centre national pour la numérisation des ressources textuelles et lexicales
CN2SV	Centre national pour la numérisation de sources visuelles
CRDO	Centre de ressources pour la description de l'oral
CRN	Centre de ressources numériques
DAF	Direction des archives de France
DC	Dublin Core (format standard de métadonnées)
EPS	Encapsulated PostScript
EXIF	Exchangeable image file format
IN2P3	Institut national de physique nucléaire et de physique des particules
LPL	laboratoire parole et langage
OAI	Open Archive Initiative
OAIS	Système ouvert d'archivage d'information (Open Archival Information System)
OLAC	Open Language Archives Community
PDF	Portable Document Format

RISC	Relais d'information sur les sciences de la cognition
SIP	Paquet d'information à verser (Submission Information Package) (terminologie OAIS)
SVG	Scalable Vector Graphics
TEI	Text Encoding Initiative
WAV	Extension utilisée pour les fichiers au format WAVE (contraction de Waveform audio format)
XML	Extensible Markup Language

12.3. Liens Internet

CINES	http://www.cines.fr/
CRDO Aix	http://crdo.fr/
CRDO Paris	http://crdo.risc.cnrs.fr/
CC IN2P3	http://cc.in2p3.fr/
RISC	http://www.risc.cnrs.fr/
TGE-Adonis	http://www.tge-adonis.fr/

12.4. Références

[REF 09]	TGE ADONIS, Mutualisation de la pérennisation et de l'accès aux données en SHS -Projet pilote sur les données orales, version 1.2, 2 février 2009. http://www.tge-adonis.fr/wiki/uploads/c/c1/ArchivageMutualise-document-synthese-v1.2.pdf
----------	--