

30/06/2008

Version 0.7

TGE ADONIS

Mutualisation de la pérennisation et de
l'accès aux données

Projet pilote sur les données orales

Table des matières

1. Objet du document.....	4
2. Les objectifs du projet.....	6
2.1. Pourquoi mutualiser ?	6
2.2. Pourquoi les données orales ?	6
2.3. Ce qu'on attend du projet pilote	7
2.4. Contrats de service	8
3. Cadre normatif général pour la conservation et pour l'accès : le Modèle de référence OAIS	9
3.1. Généralités sur le modèle OAIS	9
3.2. Le Modèle d'environnement de l'Archive	10
3.3. Le modèle d'information	11
3.3.1. Objet-information et Objet-données	11
3.3.2. Les catégories d'information	12
3.3.3. Les paquets d'information	13
3.4. Le modèle fonctionnel	14
4. Les acteurs du projet pilote	17
4.1. Le TGE Adonis	17
4.2. Les producteurs de données orales	17
4.3. Les utilisateurs de données orales.....	17
4.4. Le CRDO	17
4.5. Le CINES.....	18
4.6. Le Centre de Calcul de l'IN2P3	18
5. L'architecture fonctionnelle d'ensemble et la répartition des responsabilités et fonctionnalités entre les acteurs	20
5.1. Schéma d'ensemble	20
5.2. Répartition des responsabilités	20
5.3. Répartitions des fonctions principales	21
5.3.1. Le CRDO	21
5.3.2. Le CINES	22
5.3.3. L'IN2P3.....	23
5.4. Eléments caractéristiques des données orales.....	23
5.5. Exigences fonctionnelles	24
5.5.1. Métadonnées	24
5.5.2. Les identifiants	25
5.5.3. La vie des données.....	25
5.5.4. Le dépôt des données.....	25
5.5.5. Les droits relatifs aux données	25
5.5.6. La recherche, la sélection et l'accès aux données	26
5.5.7. Les services à valeur ajoutée	26
5.5.8. Généricité du système.....	27
5.5.9. Aide aux utilisateurs	27
5.5.10. Outils collaboratifs	27
5.6. Exigences non fonctionnelles	27
5.6.1. Performances et disponibilité du service	27
5.6.2. Sécurité	28
6. Interfaces	29
6.1. CRDO/CINES.....	29
6.2. CINES-IN2P3	29
6.3. IN2P3/CRDO.....	29
7. Acronymes et définitions	30
7.1. Glossaire	30
7.2. Abréviations	32

8. Références bibliographiques	34
9. Annexes	35
9.1. Volumétrie	35

1. Objet du document

Dans le cadre du TGE Adonis, un projet pilote consistant à développer et à mettre en exploitation une solution mutualisée de pérennisation et d'accès aux données orales numériques a été décidé.

L'objectif principal de ce projet est de valider le schéma retenu pour les données orales, tant sur le plan des fonctionnalités d'ensemble que sur le plan de la répartition des tâches et des responsabilités entre les acteurs afin de pouvoir l'étendre aux autres données des Sciences Humaines et Sociales (SHS) créées, gérées et utilisées par différents Centres de Ressources Numériques¹.

Le rapport demandé à Olof Barring du CERN² [Barring 08], dont les conclusions ont été examinées par le Conseil scientifique du TGE le 18 janvier 2008 et par le Conseil de pilotage du 24 janvier 2008, recommandait d'asseoir une solution couplant hébergement et archivage sur deux centres de calcul « lourds ». Il précisait également que compte-tenu de leurs expériences respectives, il convenait de retenir pour cela le CINES et le Centre de calcul (CC) de l'IN2P3.

Dans ce contexte, le projet pilote s'articule sur trois acteurs essentiels :

- Le CRDO, Centre de Ressources pour la Description de l'Oral, constitué d'un groupe parisien³ et d'un pôle à Aix-en Provence⁴,
- Le CINES⁵, Centre Informatique National de l'Enseignement Supérieur à Montpellier,
- Le Centre de Calcul de l'IN2P3⁶, Institut national de Physique Nucléaire et de Physique des Particules.

Le TGE ADONIS est le maître d'ouvrage du projet.

L'objet du présent document est de préciser les objectifs et le périmètre du projet pilote, de définir son cadre organisationnel, d'identifier les fonctionnalités essentielles du système distribué mis en œuvre ainsi que les rôles et responsabilités respectives des acteurs. Il servira de référence commune à ces différents acteurs qui devront, sur cette base, décrire de façon plus détaillée, les caractéristiques fonctionnelles, techniques et scientifiques de leurs rôles spécifiques et de leur infrastructure technique.

Il n'est pas exclu que certaines fonctionnalités, identifiées dans ce document, fassent l'objet d'un arbitrage par le TGE après discussion avec les partenaires du projet, afin de distinguer les fonctions indispensables au projet pilote de celles qui pourraient faire l'objet de développements ultérieurs.

Nous n'aborderons pas ici la question des solutions techniques qui seront retenues par les différents acteurs en réponse aux besoins fonctionnels. Nous n'aborderons pas non plus la question des ressources nécessaires qui font l'objet de conventions spécifiques.

¹ Les Centres de ressources numériques (CRN) ont été créés en 2006 par la Direction de l'information scientifique (DIS) du CNRS pour accompagner le passage au numérique en sciences humaines. Ils sont organisés par types de documents traités : oraux, écrits, manuscrits, géo-spatiaux, images fixes.

² Ce rapport n'a pas vocation à être mis en ligne pour un accès public. Il doit être demandé directement au TGE Adonis.

³ <http://crdo.risc.cnrs.fr/exist/crdo/>

⁴ <http://crdo.up.univ-aix.fr/index.php?langue=fr>

⁵ <http://www.cines.fr/>

⁶ http://cc.in2p3.fr/cc_accueil.php3?lang=fr

2. Les objectifs du projet

2.1. Pourquoi mutualiser ?

La mutualisation des fonctions de conservation et d'accès vise à optimiser les ressources consacrées à la recherche dans le domaine des SHS et à mieux valoriser les produits de cette recherche :

- la mutualisation des infrastructures de conservation et d'accès aux données permet des économies d'échelle et évite l'éparpillement de multiples moyens de stockage et d'accès qui nécessitent administration, gestion, maintenance, renouvellement, etc., et par conséquent dispersion et multiplication des ressources locales. Cette mutualisation permet également de disposer de compétences de haut niveau dans des domaines qui ne relèvent pas des SHS : archivage long terme de données numériques, administration et maintien en fonctionnement d'équipements informatiques (ordinateurs, moyens de stockage, réseaux) performants. Seule une masse critique d'activités permet réellement de développer ces compétences,
- la conservation à long terme des données produites dans le cadre des travaux de recherche en SHS et le maintien à long terme de l'accès à ces données doit permettre une utilisation gratuite plus large et plus durable de ces données et par conséquent constituer une meilleure valorisation des travaux qui ont permis de les produire,
- la mutualisation proposée doit aussi réduire la déperdition d'énergie au niveau des équipes locales, pour des activités qui ne relèvent pas des métiers des SHS et par conséquent une plus forte concentration sur la production et l'analyse de données et de documents ainsi que sur le développement, la mise au point et l'utilisation d'outils à valeur ajoutée spécifiques.

La mise en place d'une solution mutualisée pour la pérennisation et l'accès aux corpus oraux vise à développer une solution générique, c'est à dire une solution qui sera appliquée aux corpus oraux mais qui devra aussi être appliquée aux autres données des SHS, voire plus largement.

2.2. Pourquoi les données orales ?

Les raisons du choix des données orales comme filière expérimentale sont les suivantes :

- les données « orales » ne se réduisent pas à des enregistrements sonores : elles associent du texte (transcriptions, translittérations, annotations) et parfois des enregistrements vidéos – elles permettent donc de préfigurer partiellement la mise en place d'autres filières,
- les données déjà accessibles via le CRDO, centre de ressources numériques (CRN) créé par la DIS (Direction de l'Information Scientifique) en 2006, totalisent un nombre suffisamment important de fichiers et de métadonnées pour qu'une expérience les mettant en jeu soit réaliste,
- le domaine est marqué par une réflexion bien avancée sur les bonnes pratiques (cadres normatifs, problèmes juridiques) que synthétise l'ouvrage *Corpus oraux - Guide des bonnes pratiques* (O. Baude, Presses universitaires d'Orléans & CNRS Éditions, 2006, Paris), par ailleurs disponible en ligne⁷,
- le domaine des ressources orales est l'occasion de coopérations entre le Ministère de la Culture (MRT - Mission Recherche et Technologie et DGLFLF – Délégation Générale à la Langue Française et aux Langues de France) et le CNRS,
- une partie des usages de telles données a été précisée, en particulier à l'occasion de l'ouverture du portail « Corpus de la parole⁸ » par la DGLFLF,

⁷

http://www.dglflf.culture.gouv.fr/recherche/corpus_parole/Corpus_Oraux_GBP%202006_version_imprimee.pdf?CV=5584&type1=Ouvrage

⁸ <http://www.corpusdelaparole.culture.fr/>

- les « producteurs » de ressources sont connus, qu'il s'agisse de laboratoires CNRS ayant des programmes à long terme dans le domaine (LPL⁹, LACITO¹⁰, CLAPI¹¹ à ICAR¹², THESOC¹³ à Bases Corpus et Langage et CLLE-ERSS¹⁴ ; PFC¹⁵ à MoDyCo¹⁶ et CLLE-ERSS) ou des actions plus ponctuelles (corpus d'un chercheur ou d'un groupe de chercheurs) ou qu'il s'agisse de chercheurs ou d'équipes purement universitaires, via les actions menées en commun depuis 2004 par la DGLFLF et les deux fédérations en linguistique (TUL – Typologie et Universaux Linguistiques¹⁷, ILF – Institut de Linguistique Française¹⁸) sur un corpus de la parole et sur la participation au programme Numérisation du Ministère de la Culture¹⁹.

Il s'agit dans une grande partie des cas de données collectées par des chercheurs exerçant une mission publique dans le cadre de leurs fonctions et par conséquent il s'agit d'archives publiques dont les modalités de prise en charge, conservation et mise en valeur relèvent du Code du patrimoine, livre II sur les archives. Par conséquent, il convient, pour ces sources de données, que soit bien définie l'articulation entre d'une part la mission qui pourrait être confiée au CINES et à l'IN2P3 et celles induites par le Code du patrimoine en matière de collecte, conservation et mise en valeur pour ce qui est de l'institution responsable de ces archives au sein du ministère de la culture (services relevant de la Direction des Archives de France, voire pour certains de ces corpus, du département audiovisuel de la Bibliothèque Nationale de France. La question se pose en particulier de la délimitation entre l'archivage à des fins patrimoniales et l'archivage de données et de ressources qui sont effectivement utilisées ou utilisables à des fins de recherche.

2.3. Ce qu'on attend du projet pilote

Ce projet pilote doit permettre :

- de valider les fonctionnalités d'ensemble de la solution,
- de valider son caractère générique en dissociant ce qui est totalement applicable aux autres données des SHS de ce qui est spécifique des corpus oraux,
- de valider la répartition des tâches et des responsabilités entre les acteurs et d'identifier les sources éventuelles de difficultés,
- de valider l'infrastructure matérielle-logicielle mise en place, sur le plan des performances et de la fiabilité,
- de valider, avec la communauté des utilisateurs de corpus oraux, les services permettant la recherche, la sélection, la récupération de données ainsi que l'usage de services à valeur ajoutée propres au domaine concerné. La communauté visée est la communauté scientifique au sens large. Cette communauté dépasse le périmètre des SHS (les travaux sur l'apprentissage des langues ainsi que ceux en reconnaissance automatique de la parole font également usage de données orales),
- de disposer de premiers éléments sur les coûts de fonctionnement de la solution.

L'année 2009 pourrait à partir du bilan de cette première phase mettre en place une extension à d'autres données. *Un groupe de travail sera constitué pour suivre le projet pilote et réfléchir à sa généralisation*

⁹ <http://www.lpl.univ-aix.fr/>

¹⁰ <http://lacito.vjf.cnrs.fr/>

¹¹ <http://clapi.univ-lyon2.fr/>

¹² <http://icar.univ-lyon2.fr/>

¹³ <http://thesaurus.unice.fr/>

¹⁴ http://clle.univ-tlse2.fr/96696982/0/fiche___pagelibre/&RH=

¹⁵ <http://www.projet-pfc.net/>

¹⁶ <http://www.modyco.fr/>

¹⁷ <http://www.typologie.cnrs.fr/>

¹⁸ <http://www.ilf.cnrs.fr/>

¹⁹ http://www.culture.fr/sections/themes/culture_multimedia/sous_themes/aides-financieres/en-france/plan-national

éventuelle et aux conditions de cette dernière. Ce groupe de travail serait constitué de producteurs et d'utilisateurs de données orales qui pourraient examiner l'apport effectif de ce projet pilote à leurs besoins. Dans le cas où le projet pilote n'aboutirait pas aux résultats escomptés, le TGE Adonis aurait la responsabilité de la suite à donner (rétrocession des dépôts, etc.).

2.4. Contrats de service

Dans la perspective de la constitution d'une organisation opérationnelle et pérenne pour la conservation des données des SHS et pour l'accès à ces données, un contrat de service devra être établi entre les déposants et le TGE afin de préciser les engagements et responsabilités mutuelles. Ce contrat de service fera l'objet d'une négociation entre les déposants et Adonis. Il abordera en particulier :

- Les responsabilités d'Adonis (et des centres informatiques sur lesquels le TGE s'appuie) en matière de pérennisation des données, en matière de protection de ces données contre tout accès non autorisé, en matière de fonctionnalités d'accès et de disponibilités de service tant pour le dépôt que pour l'accès,
- Les responsabilités et engagement des déposants dans le domaine des formats de données et de métadonnées et du respect des procédures de dépôt.

3. Cadre normatif général pour la conservation et pour l'accès : le Modèle de référence OAIS

La section qui suit présente le modèle de référence OAIS qui sert de cadre normatif général pour la conservation et pour l'accès. Cette section fournit à la fois un aperçu d'ensemble et les notions précises nécessaires pour éviter les approximations.

On pourra le parcourir rapidement pour y revenir ensuite au regard du projet pilote projeté. Deux sections à la fin du document fournissent par ailleurs les définitions de ces notions et le sens des acronymes : on s'y reportera si nécessaire.

3.1. Généralités sur le modèle OAIS

Avant de rentrer plus en détail sur le projet, il est nécessaire de clarifier un minimum le vocabulaire et les concepts que nous utilisons en matière d'archivage long terme de l'information sous forme numérique.

La norme de référence fondamentale sur laquelle nous nous appuyons est le modèle de référence OAIS (Open Archival Information System) [OAIS 02]. Cette norme ISO (ISO 14721) a été élaborée par le Comité Consultatif pour les Systèmes de Données Spatiales (CCSDS), qui est une structure de normalisation commune aux agences spatiales et qui joue en même temps le rôle de comité technique au sein de l'ISO.

L'élaboration de cette norme a fait l'objet d'un travail de coopération exemplaire entre les agences spatiales principalement préoccupées par la conservation à long terme des grandes bases de données scientifiques, les grandes bibliothèques et les archives institutionnelles. En dehors du domaine spatial, cette norme est unanimement reconnue et utilisée au niveau international. En France, elle a servi de base de réflexion pour le développement de la plate-forme pilote PILAE développée pour les Archives nationales ainsi que pour le développement de la plate-forme d'archivage du CINES. Le Modèle de référence OAIS a également été défini comme la norme applicable au développement du Système de Préservation et d'Archivage Réparti (SPAR) de la BnF.

Ce Modèle de Référence définit un cadre pour la compréhension de tout ce que la préservation à long terme des données numériques peut impliquer comme principes de base, concepts, fonctions et activités. Il nous conduit à nous poser toutes les questions indispensables. Cependant, il ne propose aucune solution particulière, ni en termes d'organisation, ni en termes d'implémentation d'un système matériel et logiciel. Cette indépendance par rapport à toute implémentation et par rapport aux technologies constamment changeantes lui confère force et portée.

Nous n'en rappelons ici que quelques éléments essentiels à la compréhension du présent document. On pourra se reporter à la norme elle-même accessible librement en version anglaise ou française (voir section bibliographie).

Une Archive est définie comme une organisation chargée de conserver l'information pour permettre à une Communauté d'utilisateurs cible d'y accéder et de l'utiliser. Cette définition intègre à la fois la question de la conservation et celle de l'accès. En outre, elle implique que les données ou les documents auxquels on accède sont compréhensibles pour les utilisateurs auxquels ils sont destinés.

Il est fait une distinction claire entre :

- l'information, définie comme une connaissance pouvant être échangée.

et

- les données qui constituent une représentation formalisée de l'information, adaptée à la communication, l'interprétation ou le traitement.

Une séquence de bits, un tableau de nombres, les caractères d'une page, un enregistrement audio, etc. sont des données. Ces données sont porteuses d'une information. Il est indispensable de faire la distinction entre les deux.

3.2. Le Modèle d'environnement de l'Archive

Ce modèle est très simple, il vise simplement à délimiter les rôles et les responsabilités des acteurs.

Le **Producteur** correspond au rôle joué par les personnes (ou systèmes) qui fournissent des données à l'Archive.

Le **Management** est l'entité qui définit la mission et la politique globale de l'Archive. Le Management n'est pas concerné par le fonctionnement quotidien de l'Archive. Il ne doit pas être confondu avec l'administration au jour le jour de l'Archive qui constitue une fonction interne.

Les **Utilisateurs** sont les clients finaux (personnes ou systèmes) qui recherchent et récupèrent de l'information utile à leurs travaux.

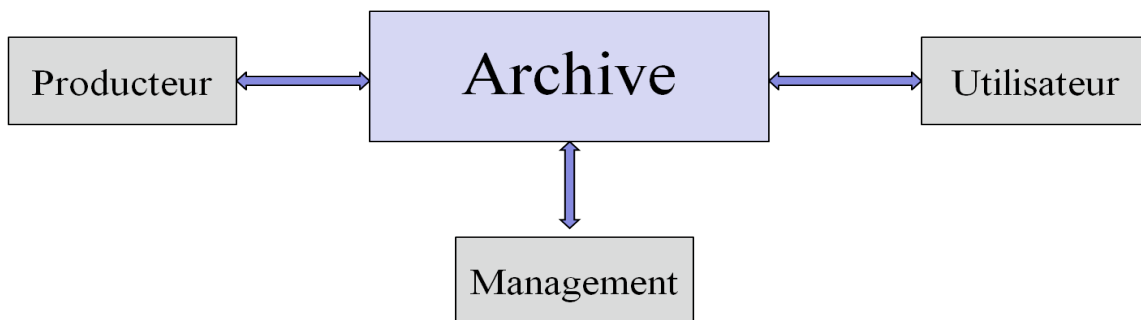


Figure 1: L'environnement de l'Archive

Par rapport à ce schéma, le rôle, les responsabilités et les fonctions des différents acteurs du projet seront explicités dans les chapitres 4 et 5 ci-après.

3.3. Le modèle d'information

3.3.1. Objet-information et Objet-données

Le Modèle d'Information définit un certain nombre de catégories d'information, permettant de distinguer l'objet cible de la pérennisation, des autres types d'information servant à décrire et assurer la compréhension de cet objet sur le long terme.

En pratique ce modèle est construit à partir du concept d'**Objet-information**, qui se décline en plusieurs catégories d'Information. Ces catégories sont regroupées en paquets cohérents permettant la manipulation, la gestion et la circulation de l'information au sein de l'Archive et avec son environnement extérieur.

L'**Objet-données** est défini comme un ensemble de séquences de bits. Comme nous l'avons indiqué précédemment, l'information est représentée par des données. A cet Objet-données, nous ferons donc

correspondre un **Objet-information** modélisant le contenu informationnel de l'Objet-données

L'interprétation des données ne peut nous conduire à l'information représentée par ces données qu'à la seule condition de disposer d'une catégorie d'information particulière appelée '**Information de représentation**'. C'est ce que montre la figure 2 ci-après que nous allons illustrer par un exemple très simple permettant d'éclairer cette relation entre données et information.

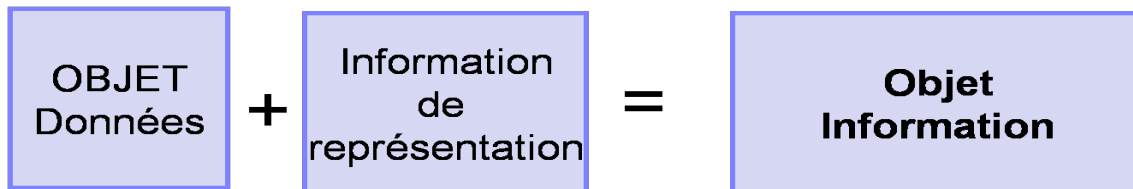


Figure 2: Des données à l'information contenue dans ces données

Prenons le cas d'un objet 'Données' constitué par une chaîne de 128 bits à l'intérieur d'un fichier. Supposons que cette chaîne soit un texte codé en conformité avec la norme ISO 8859-1 (Technologies de l'Information – Jeu de caractères graphiques codés sur un seul octet : Alphabet latin N°1). Connaissant cette information nouvelle, à savoir que ces bits doivent être interprétés en appliquant la norme ISO 8959-1, il est alors possible d'interpréter les données comme une séquence de caractères. Ce processus a permis de transformer l'objet-données (la chaîne de bits), à l'aide de la norme ISO 8859-1 (l'information de représentation) en un Objet- information plus intelligible que la chaîne de bits originale. On peut donc souligner ici que pour préserver l'objet-information, il est indispensable de préserver la norme ISO 8859-1 ainsi que l'association qui existe entre la chaîne de bits et la norme ISO 8859-1.

L'Information de Représentation peut évidemment être infiniment plus détaillée que la norme ISO 8859-1 et l'objet-données peut être beaucoup plus complexe qu'une séquence de caractères. Ceci est particulièrement vrai pour les données orales pour lesquelles nous serons amenés à manipuler des objets plus complexes.

3.3.2. Les catégories d'information

Le '**Contenu d'information**' est un ensemble d'information qui constituent l'objet principal de la pérennisation. Il sera constitué des données (Objet-données) et de l'Information de représentation associée.

L'Information de pérennisation est l'information nécessaire à une bonne conservation du Contenu d'information. Elle est décomposée en quatre catégories : informations de provenance, d'identification, d'intégrité et de contexte, dont voici une brève description :

- L'Information de provenance documente l'historique du Contenu d'information. Cette information renseigne sur l'origine du Contenu d'information, sur toute modification intervenue depuis sa création,
- L'information de contexte décrit les liens entre un Contenu d'information et son environnement. Elle inclut entre autres les raisons de la création de ce Contenu d'information et son rapport avec d'autres Objets-contenu d'information. Par exemple, elle peut expliquer pourquoi le Contenu d'information a été produit, et inclure une description de la façon dont ce Contenu est relié à un autre Objet-contenu d'information existant,

- L'information d'identification identifie, et si nécessaire, décrit le ou les mécanismes d'attribution des identificateurs au Contenu d'information. Elle inclut aussi les identificateurs qui permettent à un système externe de se référer sans équivoque à un Contenu d'information particulier. Exemple : un ISBN (International Standard Book Number),
- L'Information d'intégrité décrit des mécanismes et des clés d'authentification garantissant que le Contenu d'information n'a pas subi de modification sans que celle-ci ait été tracée. Par exemple, l'empreinte du fichier calculée à partir d'un algorithme de hachage.

Nous verrons plus loin que le Contenu d'information et l'information de pérennisation sont réunis au sein d'un conteneur conceptuel appelé Paquet d'informations archivés (AIP - Archival Information Package).

L'Information d'empaquetage est l'information permettant de relier les composants d'un paquet d'informations. Par exemple les informations de volume et de répertoire dans un CD-ROM conforme à la norme ISO 9660, permettant d'accéder aux fichiers du Contenu d'information et de l'Information de pérennisation. Elle peut s'appuyer sur le système de gestion de fichiers, sur la structure des répertoires, elle peut utiliser des systèmes de pointeurs et des méta-langages génériques tels que XML.

L'Information Descriptive est l'ensemble d'informations, constitué principalement de Descriptions de paquet. L'Information Descriptive est l'information qui sera utilisée dans ce processus de recherche et d'évaluation de l'intérêt potentiel d'une information archivée par rapport à un objectif de l'utilisateur. Cette Information descriptive contient les données d'entrée des Outils d'accès. Elle est le plus souvent organisée au sein d'une base de données.

La figure 3 ci-après présente sous la forme d'un diagramme UML, les différentes catégories d'information décrites dans le Modèle OAIS.

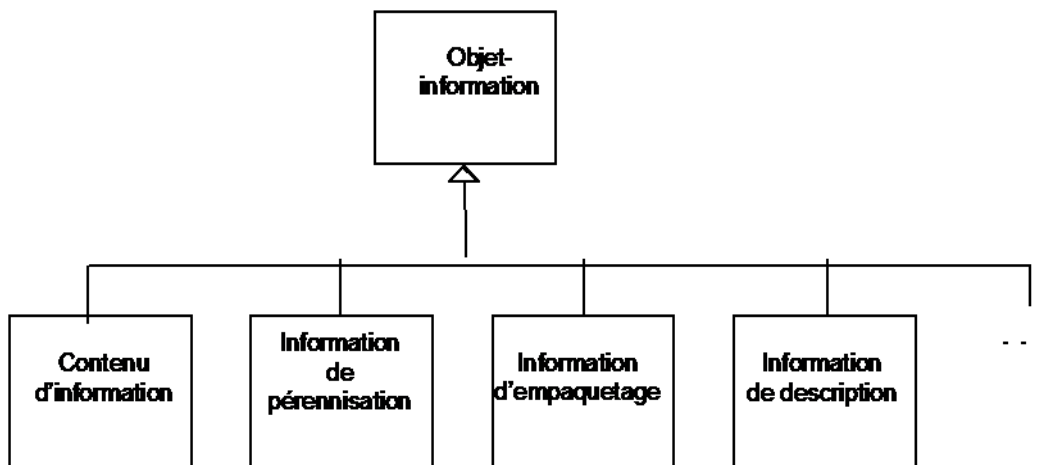


Figure 3: Taxinomie de l'Objet-information

3.3.3. Les paquets d'information

Le **Paquet d'informations** est un concept essentiel du Modèle d'Information de l'OAIS.

Le Paquet d'Informations est l'association du Contenu d'information et de son information de pérennisation destinée à faciliter la conservation du Contenu d'information. A ce paquet d'informations sont associées :

- une Information d'empaquetage utilisée pour circonscrire et identifier le Contenu d'information et

- l'information de pérennisation associée,
- une information descriptive du paquet.

La figure 4 ci-après présente une modélisation UML du paquet d'information :

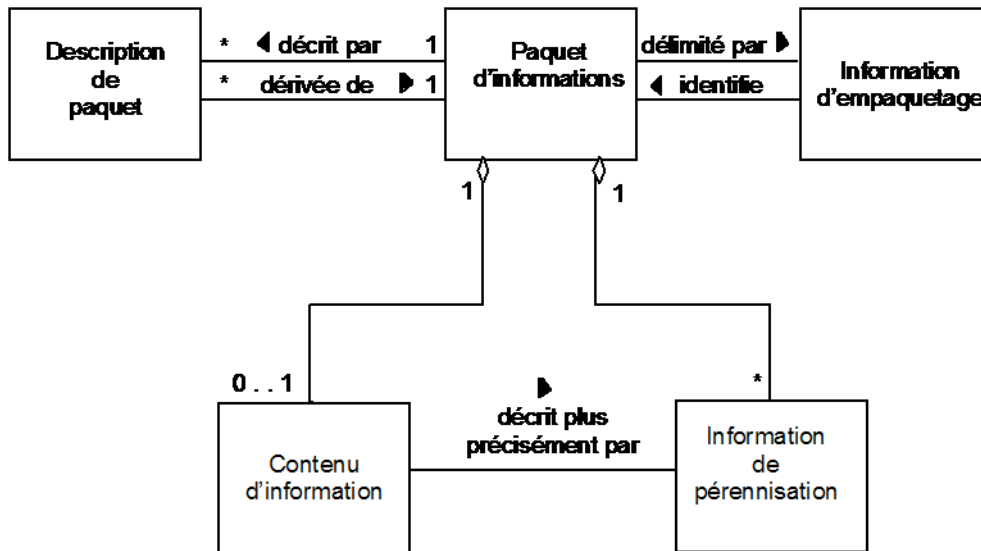


Figure 4: Le Paquet d'information

Il existe trois variantes du Paquet d'informations, en fonction du contexte dans lequel ce paquet est utilisé dans l'environnement de l'Archive :

- Le **Paquet d'Information à verser (SIP, Submission Information Package)** est le Paquet d'informations livré par le Producteur à l'Archive pour l'élaboration d'un ou plusieurs Paquets d'informations archivés. Sa forme et son contenu détaillé sont négociés entre le Producteur et l'Archive.
- Au sein de l'OAIS, un ou plusieurs SIP sont transformés en un ou plusieurs **Paquets d'informations Archivés (AIP – Archival Information package)** afin d'être préservés. Plusieurs SIP peuvent être utilisés pour créer un seul AIP et inversement un seul SIP peut conduire à la création de plusieurs AIP.
- En réponse à une requête, l'Archive fournit tout ou partie d'un AIP à un Utilisateur sous la forme d'un **Paquet d'informations Diffusé (DIP – Dissemination Information Package)**. Ce paquet provient d'un ou de plusieurs Paquets d'informations archivés.

Ce concept de paquet a en fait de nombreuses applications pratiques. Un grand nombre d'Archives s'appuient sur des normes d'empaquetage et sur des outils logiciels associés pour les opérations de transfert de ou vers l'environnement extérieur (SIP et DIP) mais aussi pour la création effective et concrète de paquets d'information archivés.

3.4. Le modèle fonctionnel

Six entités fonctionnelles principales ont été définies au sein de l'Archive :

L'entité « **Entrées** » assure les fonctions relatives à l'acceptation des Paquets d'informations à verser provenant des Producteurs et à la préparation de leur contenu en vue du stockage et de la gestion des données au sein de l'Archive.

L'entité « **Stockage** » assure les fonctions relatives au stockage, à la maintenance et à la récupération des Paquets d'information archivés.

L'entité « **Gestion de données** » assure les fonctions relatives à l'enrichissement, la conservation et l'accès à l'Information de description (qui identifie et documente les fonds de l'Archive) et aux données administratives utilisées pour gérer l'Archive.

L'entité « **Administration** » regroupe les fonctions permettant la supervision continue du fonctionnement des autres entités de l'Archive.

L'entité « **Planification de la pérennisation** » assure les fonctions relatives à la surveillance de l'environnement de l'Archive et à la production de recommandations visant à ce que les informations archivées restent accessibles et compréhensibles sur le long terme pour la Communauté d'utilisateurs cible, même si l'environnement informatique d'origine devient obsolète.

L'entité « **Accès** » assure les fonctions qui aident l'Utilisateur à déterminer si une information existe ou non dans l'Archive, à trouver sa description, son emplacement si elle est disponible, et à demander et recevoir des produits d'information.

Le modèle fonctionnel de l'Archive, replacé dans son environnement est présenté sur la figure 5 ci-après.

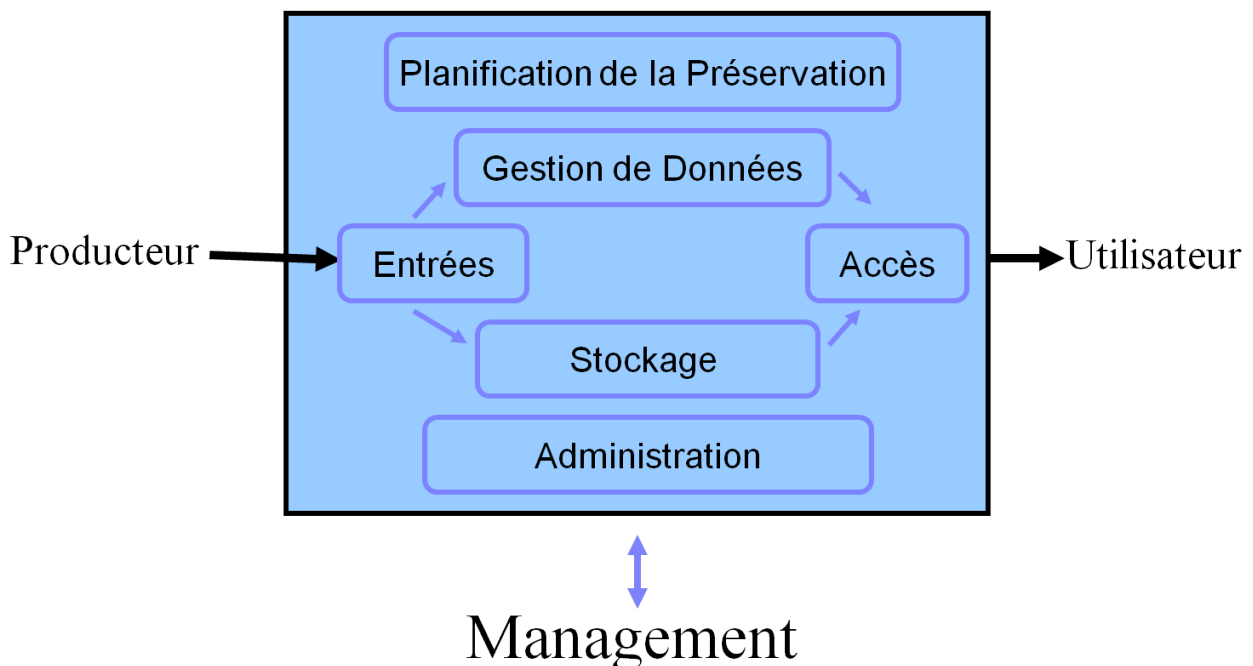


Figure 5: Le modèle fonctionnel de l'Archive

Dans le Modèle de Référence, les entités fonctionnelles définies ci-avant sont décomposées en sous-fonctions permettant de définir plus précisément l'enchaînement des différentes opérations.

4. Les acteurs du projet pilote

4.1. Le TGE Adonis

Le TGE Adonis est le maître d'ouvrage du projet. Ce projet s'appuyant sur plusieurs acteurs indépendants, c'est au TGE qu'il revient de piloter le projet dans son ensemble et d'assurer la coordination entre les acteurs et si besoin, d'arbitrer les désaccords. C'est au TGE également qu'il appartient d'anticiper la préparation de l'extension du schéma retenu pour le projet à d'autres corpus des SHS.

4.2. Les producteurs de données orales

Les producteurs de corpus peuvent être institutionnels (laboratoires), ils peuvent être des regroupements d'équipes (réunies par un projet ANR par exemple), des chercheurs isolés.

On peut distinguer plusieurs situations :

1. les enregistrements (et éventuellement les transcriptions) qui ont été effectués dans le passé, mais qui peuvent servir encore pour des recherches actuelles ;
2. les laboratoires qui ont une activité récurrente de constitution et exploitation de données orales ;
3. les projets sur quelques années qui supposent la constitution et l'exploitation de données orales.

Dans le premier cas, il s'agit de numériser des enregistrements analogiques, éventuellement de scanner des transcriptions et des annotations sur support papier et de documenter (métadonnées) l'ensemble. Ce sont parfois des fonds en déshérence, ce qui ne facilite pas la gestion des droits : les chercheurs les ayant constitués ne sont plus joignables.

Dans les deux autres cas, les « facteurs » d'enregistrements et d'annotations peuvent souhaiter un embargo limité dans le temps de l'accès à ces ressources (pour permettre l'exploitation, la publication). Il en découle la nécessité de gérer l'accréditation et l'authentification de l'accès à ces données. Par contre, les données et métadonnées ne sont pas pour autant d'emblée dans les formats normatifs du domaine. Cette normalisation peut ne pas avoir été prévue en amont, ce qui fait qu'elle devient difficile à réaliser, faute de financement.

4.3. Les utilisateurs de données orales

Les utilisateurs sont des équipes ou des chercheurs utilisant des corpus pour leurs travaux. Ils ne se limitent pas à la communauté des SHS. Ils ne se limitent pas non plus à la communauté française.

4.4. Le CRDO

Dans le domaine des corpus oraux, c'est le CRDO qui concentre et coordonne la compétence et l'expertise métier. Cette compétence métier sera mise à profil dans deux domaines essentiels :

- Celui de la connaissance des données orales et les métadonnées qui les décrivent. Le CRDO jouera à ce titre, un rôle essentiel de préparation des données et des métadonnées en vue de leur archivage,
- Celui de la connaissance des outils spécifiques aux données orales, tant pour la recherche d'information que pour l'analyse de cette information.

4.5. Le CINES

Dans le domaine du calcul numérique intensif, le CINES offre aux laboratoires la possibilité de paralléliser et d'exploiter leurs codes scientifiques. De nombreuses disciplines scientifiques utilisent les équipements du Centre pour la résolution de problèmes qui exigent des puissances de calcul extrêmes et de grandes capacités de mémoire. Le CINES dispose de moyens informatiques, d'équipements de stockage et de connexions réseaux performants ainsi qu'une expertise reconnue en matière de maintien en fonctionnement opérationnel d'un centre de calcul.

Le Ministère de l'Enseignement supérieur et de la Recherche, ministère de tutelle du CINES, lui a confié également une mission nationale d'archivage pérenne de documents numériques du patrimoine scientifique. Dans le cadre de cette mission, il a développé une compétence solide dans ce domaine. Il a mené depuis cinq ans, un ensemble de travaux et de développement pour la prise en charge de l'archivage numérique des thèses des universités françaises²⁰ et pour l'archivage des revues scientifiques en sciences humaines et sociales du portail PERSEE²¹.

Les compétences du CINES en matière d'archivage couvrent l'ensemble des entités fonctionnelles du modèle OAIS mais elles sont particulièrement concentrées sur les fonctions d'entrées (mise en place d'un dispositif de transfert des données et des métadonnées, validation de la conformité des objets numériques reçus par rapport aux exigences relatives aux formats et aux métadonnées, création des paquets d'information à archiver) et sur les fonctions de stockage (organisation du stockage, suivi, gestion, surveillance, contrôle des empreintes...).

La mission du CINES dans le domaine de l'archivage a récemment été confortée par la tutelle. La lettre de mission des Directeurs de la DGRI (Direction générale de la recherche et de l'innovation) et de la DGES (Direction générale de l'enseignement supérieur) au directeur du CINES, du 12 février 2008, précise : « Concernant l'archivage pérenne, les enjeux pour le CINES sont l'acquisition d'une nouvelle compétence métier. C'est avec ses partenaires (ministères et universités) que le CINES devra également faire progresser l'intégration de la chaîne fonctionnelle d'archivage qui remonte aux auteurs et responsables des documents ».

4.6. Le Centre de Calcul de l'IN2P3

Le CC IN2P3 dispose également de moyens informatiques, d'équipement de stockage et de connexions réseaux très performants. Il a développé une expertise dans plusieurs domaines, en particulier la conception, l'installation et l'exploitation des fermes de stations pour le traitement des données, la conception et la mise en œuvre d'une architecture pour le stockage massif de données et la mise en place de l'infrastructure de transport des données (réseaux locaux et étendus à très haut débit). Le CC-IN2P3 a également ouvert ses portes aux astrophysiciens et aux biologistes²². Le CC-IN2P3 a également développé une expertise dans les technologies de grille informatique (*grid computing*) et est devenu aujourd'hui un acteur majeur du dispositif en France.

Le CC IN2P3 héberge différents sites comme celui du Centre National pour la Numérisation des Sources Visuelles (CN2SV²³) ou encore le serveur HAL²⁴ (Hyper Articles en Ligne) du CCSD²⁵ (Centre pour la Communication Scientifique Directe). HAL permet de déposer et de rendre publics des documents scientifiques de toutes les disciplines.

²⁰ http://cct.cnes.fr/cct24/public/2007/seminaires/archivage_theses_CINES/seminaire_pmi_archivage_theses_cines.pdf

²¹ <http://www.persee.fr/>

²² <http://cc.in2p3.fr/rubrique131.html>

²³ <http://www.cn2sv.fr/spip.php?article110>

²⁴ <http://hal.archives-ouvertes.fr/>

²⁵ <http://www.ccsd.cnrs.fr/>

5. L'architecture fonctionnelle d'ensemble et la répartition des responsabilités et fonctionnalités entre les acteurs

5.1. Schéma d'ensemble

La figure 6 présentée ici constitue la traduction, dans le cadre du projet pilote du TGE Adonis sur les corpus oraux, du modèle fonctionnel de la norme OAIS.

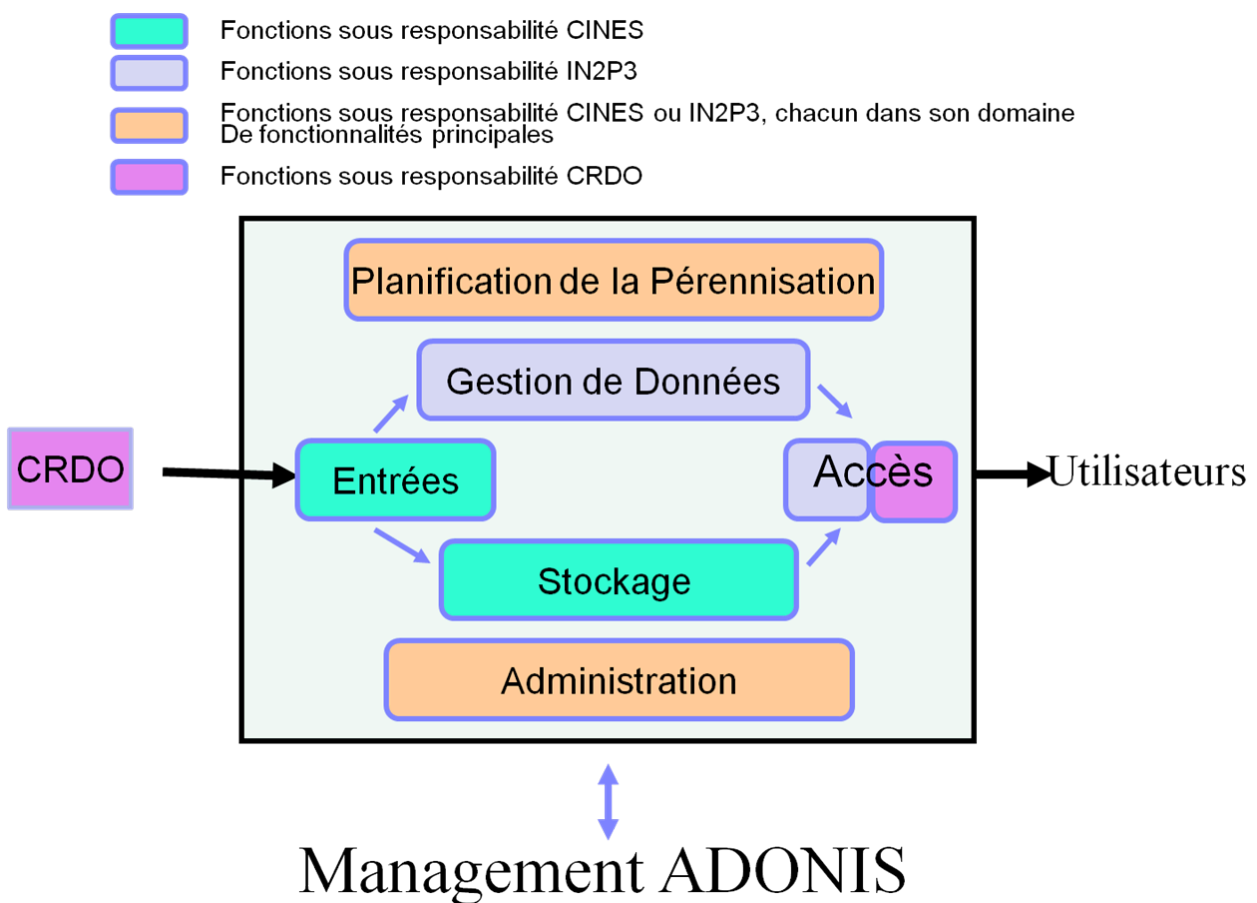


Figure 6: schéma fonctionnel du projet pilote

5.2. Répartition des responsabilités

C'est le TGE ADONIS qui porte la responsabilité d'ensemble de la définition, de l'organisation et de la conduite du projet. Le TGE assure donc le rôle de chef de projet : il définit les fonctionnalités d'ensemble du

système et le partage de ces fonctionnalités entre les acteurs, il définit les tâches à entreprendre et l'ordonnancement de ces tâches conformément au calendrier objectif, il gère les actions du projet, maintient à jour les documents de référence du projet, assure la coordination entre le CRDO, le CINES et l'IN2P3, au travers d'un comité de pilotage, définit les critères de validation des résultats obtenus par le projet, organise les tests d'ensemble en relation avec la communauté des corpus oraux et le CRDO .

Le CRDO est responsable de la collecte des objets, du contrôle de leur validité sémantique et de leur conformité aux exigences d'archivage en matière de formats de données et de métadonnées. Le CRDO est également responsable des services à valeur ajoutée métier.

Le CINES est responsable de la collecte des objets à archiver et du stockage pérenne de ces objets

L'IN2P3 assure la gestion des métadonnées descriptive et de mise en œuvre des fonctions d'accès aux corpus.

5.3. Répartitions des fonctions principales

Cette section définit plus en détail les fonctions induites par les responsabilités respectives des acteurs. Le TGE ADONIS n'intervient pas directement dans le processus fonctionnel de l'archivage et la diffusion des corpus oraux.

5.3.1. Le CRDO

C'est le CRDO qui porte l'expertise métier. Il intervient à la fois sur la préparation des données en vue de leur archivage et sur les services d'accès à ces données.

Préparation des données en vue de leur transfert à l'Archive :

Le CRDO sert de point d'entrée vers les producteurs de ressources, il opère un premier niveau de validation sur les données et métadonnées et transmet au CINES. Il assure les fonctions suivantes :

- collecte des données orales auprès des producteurs (en pratique, le projet pilote sera basé pour l'essentiel sur des corpus existants),
- contrôle de ces données,
- transformation si nécessaire vers un format d'archivage pérenne,
- collecte des métadonnées, validation de leur contenu, mise au format de métadonnées retenu pour le projet,
- constitution de paquets d'information associant données et métadonnées conformément à l'interface définie avec le CINES,
- transfert des données et métadonnées au CINES pour conservation sous forme de paquets d'information,
- corrections des paquets d'information rejetés par le CINES pour anomalie.

Une partie des producteurs n'est pas forcément en mesure de produire des données et métadonnées au format d'archivage requis. La mise en conformité des données et métadonnées par rapport aux standards retenus devra, au moins dans un premier temps, revenir au CRDO. Le CRDO pourra dans ce domaine capitaliser l'expérience, les méthodes et les outils. Dans l'avenir, il conviendra cependant d'apporter un support aux producteurs de données de façon à ce que les données soient chaque fois que cela est possible, produites directement dans un format d'archivage.

Afin de faciliter le travail préparatoire du CRDO dans la perspective des dépôts, un espace de travail

collaboratif sera mis en place afin de faciliter la validation de ces dépôts avant leur transfert au CINES.

Interfaces d'accès aux données et services à valeur ajoutée :

Un certain nombre d'interfaces d'accès aux données orales ont été développées. Elles devront en partie servir de modèle aux interfaces futures développées dans le cadre du projet pilote. La réutilisation de tout ou partie de ces interfaces devra être étudiée.

Différents outils ont été développés dans le cadre des activités du CRDO. Ces outils concernent l'aide à la constitution des annotations, l'aide à la sélection et à l'accès aux données orales, etc. Un inventaire des outils qu'il conviendrait de mettre en œuvre dans le cadre du projet pilote sera dressé. Certains seront installés sur le site de l'IN2P3 pour utilisation directe dans le processus d'accès aux données, d'autres devront pouvoir être utilisés de façon indirecte via un service Web. Dans ce dernier cas, ces outils seront exécutés sur le site du laboratoire développeur.

5.3.2. Le CINES

Il est principalement en charge des fonctions de collecte et validation des données à archiver, de constitution des paquets d'information archivés, de conservation pérenne de ces paquets et de transmission d'une copie des données et métadonnées pour l'accès.

Plus précisément, le CINES :

- met en place une interface de versement permettant le transfert sécurisé des objets depuis le CRDO (standardisation du mode de versement des données),
- opère la validation technique des données et des métadonnées (conformité des métadonnées aux schémas et conformité des données aux formats d'archivage). Le CINES n'effectue pas de transformation de format,
- retourne au CRDO les paquets invalides,
- complète les métadonnées en vue de l'archivage (un identifiant unique par objet, empreintes numériques, métadonnées techniques...). Remarquons qu'un identifiant affecté par le CINES ne préjuge pas de l'existence d'un autre identifiant relevant d'un consensus métier.
- crée les paquets d'information archivés,
- stocke ces paquets sur le site du CINES et assure leur conservation : surveillance des supports, migrations, sécurité des données, contrôles d'intégrité, traçabilité des opérations,
- organise un stockage de secours sur un site distant ainsi que les fonctionnalités de transfert vers ce site et de récupération à partir de ce site,
- transmet les paquets d'information archivés à l'IN2P3 pour la mise en œuvre des fonctions de gestion des données et d'accès aux utilisateurs finaux,
- produit un ensemble de statistiques sur les données archivées (nombre de services producteurs concernés, applications productrices de données concernées, nombre total de transmissions par an, volume versé par an).

5.3.3. L'IN2P3

L'IN2P3 assure principalement les fonctions de gestion des données et d'accès à ces données par les utilisateurs finaux. Ces fonctions impliquent :

- l'organisation des métadonnées descriptives et la mise en place d'un moteur de recherche générique,

- une interface graphique accessible sur Internet via un navigateur permettant la récupération de données au format de diffusion en http (download) et en rtsp (streaming),
- une interface OAI pour le moissonnage,
- l'hébergement d'un portail Web permettant au CRDO de donner toutes les informations utiles,
- la gestion des droits d'accès aux données et aux métadonnées : les utilisateurs accèdent aux différents services de recherche, sélection, récupération en fonction des droits dont ils disposent,
- la conversion des données du format d'archivage vers les formats de diffusion demandés par l'utilisateur,
- l'appel à des services à valeur ajoutée locaux ou distants,
- la production de statistiques (nombre de commandes de sorties d'archives reçues par an) typologie

5.4. Eléments caractéristiques des données orales

Il s'agit, dans cette section, de présenter les caractéristiques spécifiques des données orales qui devront être prises en compte par le projet.

Les données orales sont d'abord constituées d'enregistrements accompagnés de métadonnées. Ces enregistrements sont le plus souvent des enregistrements sonores mais pas exclusivement (cas de la langue des signes par exemple). Ils sont généralement sous la forme de fichiers au format Wave mais on pourra trouver d'autres formats, du MP3, de la vidéo. La liste exacte des formats des enregistrements acceptables pour l'archivage long terme sera établie entre le CINES et le CRDO.

Ces données orales peuvent exister en plusieurs versions (enregistrement brut, enregistrement traité...).

A ces enregistrements, sont le plus souvent associés des données complémentaires appelées annotations. Ces annotations peuvent être des transcriptions phonétiques, orthographiques, (le plus souvent alignées temporellement), elles peuvent être le résultat de tout type d'analyse effectuée sur les enregistrements. Ces annotations constituent des objets de données accompagnés de leurs métadonnées. A un enregistrement peut correspondre 0, 1 ou plusieurs annotations qui peuvent être déposées dans l'archive à des moments différents.

Le lien entre un fichier d'annotation et l'enregistrement qu'il décrit est établi par l'intermédiaire des métadonnées :

Les annotations peuvent avoir plusieurs formes différentes :

- fichier conforme à une DTD ou un schéma XML. Il n'y a pas de schéma normalisé, mais il existe un nombre limité de schémas et de DTD utilisés,
- fichiers texte structuré
- fac-similé sous forme d'images créées à partir une annotation manuscrites,
- ...

5.5. Exigences fonctionnelles

5.5.1. Métadonnées

Les métadonnées utilisées par le CRDO sont les métadonnées standard définies par OLAC²⁶ (Open Language Archives Community). Elles sont basées sur les métadonnées Dublin Core :

- elles utilisent les 15 éléments de base du Dublin Core : title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, rights ainsi que la quarantaine d'étiquettes du Dublin-Core qualifié,
- certains de ces éléments font l'objet d'une définition plus précise de la part d'OLAC pour répondre aux besoins spécifiques des données orales avec un vocabulaire contrôlé (mécanisme du Dublin Core qualifié)

Les métadonnées des enregistrements référencent chaque annotation par une relation DC "Is Required By ". Inversement, les métadonnées des annotations pointent sur l'enregistrement par une relation DC "requires".

Il s'ensuit que chaque versement d'une nouvelle annotation (et de ses métadonnées) implique une mise à jour des métadonnées de l'enregistrement concerné.

D'autres types de relations sont également pris en compte : filiations entre plusieurs transcriptions, relations de tout à partie,...

L'organisation des métadonnées est également conforme aux exigences du protocole OAI

Les métadonnées pour l'archivage devront respecter les exigences de l'archivage et prendre en compte toutes les spécificités des données orales.

La structure des métadonnées pour l'archivage devra permettre l'insertion d'autres données des SHS ayant leurs propres spécificités.

Il devra être possible de gérer la notion de collection, et de définir les métadonnées communes à tous les objets de la collection, chaque objet ayant quelques métadonnées spécifiques.

Il reste à proposer une terminologie claire pour les producteurs et les utilisateurs qui distingue :

- la ressource atomique : (partie de) transcription ou enregistrement ;
- la collection qui regroupe des ressources atomiques en fonction du contexte de leur production (dans le cadre d'une même enquête, par exemple) ;
- la base de données orales, qui regroupe ressources atomiques et collections ;
- le corpus, qui est l'« assemblage » pour un projet de recherche précis de ressources atomiques et éventuellement de collections.

Les métadonnées doivent permettre de faire apparaître clairement les différents contributeurs qui ont participé à la création de la ressource (cf. métadonnées OLAC sur ce point : annotator, author, compiler, consultant, data-inputter, depositor, developer, editor, illustrator, interpreter, interviewer, participant, performer, photographer, recorder, researcher, research_participant, responder, signer, singer, speaker, sponsor, transcriber, translator). Ces indications fines permettront en particulier de « rendre à César ce qui appartient à César » au moment de l'accès aux données, c'est-à-dire d'indiquer les « crédits » effectifs. C'est particulièrement important pour que les apports de chacun soient reconnus et qu'une solution mutualisée soit perçue comme complémentaire des laboratoires et projets existants.

²⁶ <http://www.language-archives.org/>

5.5.2. Les identifiants

Les identifiants utilisés actuellement sont des identifiants OAI non pérennes ou des URL dont la pérennité est toujours problématique.

Un système d'identifiants pérennes sera défini (analogue au DOI²⁷ – Document Object Identifier – dans le domaine des publications). Il devra pouvoir être réutilisé pour les autres domaines des SHS.

5.5.3. La vie des données

Un utilisateur doit pouvoir demander l'ensemble des versions d'un même enregistrement : on peut distinguer l'identifiant unique d'un objet d'un identifiant logique qui serait commun à toutes les versions de cet objet.

5.5.4. Le dépôt des données

Le CINES proposera une procédure de dépôt. Elle devra permettre au déposant de mettre en œuvre des dépôts automatiques (dépôts en nombre gérés par une procédure logicielle) et des dépôts manuels.

Les modalités de dépôt seront précisées dans un document d'interface CRDO-CINES (voir chapitre 6).

5.5.5. Les droits relatifs aux données

Les droits de propriété intellectuelle sont explicités dans l'attribut « right » des métadonnées.

Les droits d'accès <accessRights> et la licence d'utilisation <license> sont explicités également explicités dans les métadonnées. A ce niveau, il est simplement précisé si les données sont en accès ouvert ou restreint. Ceci implique cependant une mise à jour de ces métadonnées de droit lorsque la ressource passe de l'état « à accès restreint » à l'état « accès ouvert »

Les droits d'accès aux enregistrements et de chacune des annotations doivent être gérés de façon indépendante : tel enregistrement et telle annotation pourront être ouverts alors que telle autre annotation portant sur le même enregistrement pourra être à accès restreint.

Une protection de l'accès aux données par login et mot de passe est suffisante pour satisfaire les restrictions d'accès.

Dans le cas d'un accès restreint, la liste du ou des groupes utilisés à consulter ces données et la liste des membres de ces groupes sera gérée indépendamment des métadonnées.

5.5.6. La recherche, la sélection et l'accès aux données

Plusieurs modes de recherche devront pouvoir être proposés :

- recherche basée sur une partie des champs des métadonnées,
- recherche textuelle dans les métadonnées,
- recherche textuelle sur les transcriptions orthographiques ou phonétiques

Ceci implique qu'au niveau de la recherche de ressources, on gère des objets logiques regroupant l'enregistrement et les annotations associées.

La récupération effective des données implique :

- de se conformer aux droits d'accès qui peuvent être distincts pour les enregistrements et les annotations,

²⁷ <http://www.doi.org/>

- d'approuver un texte de licence d'utilisation propre aux données que l'on veut récupérer.

L'interface d'accès devra permettre le multilinguisme.

Elle devra permettre également l'accès à un certain nombre d'informations documentaires générales associées aux fonds.

Enfin, le système devra permettre le moissonnage OAI.

5.5.7. Les services à valeur ajoutée

Outils et services installés sur le site de l'IN2P3 :

Différents outils d'aide à la sélection seront mis à la disposition des utilisateurs. Ces outils doivent permettre aux utilisateurs d'évaluer plus rapidement l'intérêt de telle ou telle données par rapport à ses propres travaux et faciliter la récupération des informations utiles.

Sont identifiés pour l'instant :

- la diffusion en mode streaming en format dégradé pour la vidéo et l'audio,
- le découpage à la volée permettant de ne récupérer que la partie d'un enregistrement située entre deux temps donnés et non la totalité de cet enregistrement.

La liste des outils est indicative et devra pouvoir évoluer au cours de temps.

Outils et services accessibles par Web service :

A définir

5.5.8. Généricité du système

Le système d'ensemble a une vocation générique. Il devra naturellement pouvoir prendre en compte les spécificités de chaque domaine des SHS.

Dans le cahier des charges propres à chacune des parties relevant du CINES d'une part et de l'IN2P3 d'autre part, on devra spécifier ce qui est générique et ce qui est spécifique (pour les données orales dans cette phase du projet).

5.5.9. Aide aux utilisateurs

A priori l'aide sur les ressources et leur utilisation sera assurée par le CRDO pour tout ce qui concerne la compréhension des contenus. En même temps, chaque fois que c'est possible (c'est-à-dire quand il s'agit de données liées à des projets de recherche « vivants »), le lien sera établi et maintenu avec les producteurs de la ressource, de manière à ce que l'information fournie soit de première main.

Une aide aux utilisateurs pour tout problème lié au fonctionnement du système d'accès aux données devra aussi être envisagée à l'IN2P3.

5.5.10. Outils collaboratifs

Un espace de travail collaboratif sera mis en place tant pour faciliter l'élaboration des annotations (outil de concordancier...) que pour la préparation des dépôts.

Ce point sera développé ultérieurement. On peut souhaiter en particulier des espaces de travail destinés à des groupes-projets utilisant un ensemble déterminé de ressources et facilitant la présentation publique des résultats de la recherche.

5.6. Exigences non fonctionnelles

5.6.1. Performances et disponibilité du service

Ces points seront définis ultérieurement. Ils concernent :

- Les performances et la disponibilité au niveau des versements ;
- Les performances et la disponibilité au niveau de l'accès aux données par les utilisateurs (interface IN2P3)

5.6.2. Sécurité

Le cahier des charges du CINES pour Adonis devra préciser les exigences de sécurité mises en place au niveau des dépôts.

Le document d'interface CINES-IN2P3 explicitera les exigences de sécurité relatives à tous les transferts d'information entre les deux centres.

Le cahier des charge de l'IN2P3 pour Adonis devra préciser les exigences de sécurité du système au niveau des fonctions de recherche, de sélection et de récupération de données ainsi qu'au niveau des différents outils. Une gestion des groupes d'utilisateurs autorisés à consulter et récupérer telle ou telle donnée protégée et une identification par Login et mot de passe seront mises en place.

6. Interfaces

Les spécifications techniques d'interfaces CRDO-CINES, CRDO-IN2P3, CINES-IN2P3 seront établies au cas par cas par les partenaires concernés, avec le support du TGE. Toute difficulté sera remontée au niveau du TGE et devra être résolue à ce niveau.

La question d'une double interface (français et anglais) devra être tranchée.

6.1. CRDO/CINES

Les interfaces entre le CRDO et le CINES feront l'objet d'une spécification d'interface approuvée par les deux parties. Cette spécification d'interface précisera :

- les formats de données acceptables dans la perspective de conservation à long terme. A titre indicatif, les formats souhaités actuellement par le CRDO comprennent notamment :
 - Audio : WAV ou BWF, codage PCM, ≥ 44.1 kHz, ≥ 16 bits mono ou stéréo
 - Vidéo : DV
 - Annotations : XML, texte seul ASCII, texte seul Unicode, PdF (en tant que conteneur pour des images (scans de manuscrits)). Certains fichiers livrés en formats "texte seul" (Praat, CHAT, Shoebox) doivent être XMLisés au moment du pré-versement. L'automatisation de ces conversions est envisageable.
- les formats de métadonnées,
- les procédures et protocoles de dépôt,
- etc.

6.2. CINES-IN2P3

Les interfaces entre le CINES et l'IN2P3 feront l'objet d'une spécification établie en commun par les deux centres. Ces interfaces n'ont pas d'impact direct sur les déposants et sur les utilisateurs externes.

6.3. IN2P3/CRDO

Les interfaces entre le CRDO et l'IN2P3 feront l'objet d'une spécification d'interface approuvée par les deux parties. Cette spécification d'interface précisera :

- le mode de gestion des droits d'accès aux données (accréditation, authentification) et le rôle du CRDO dans cette gestion,
- les formats de diffusion des données et des annotations,
- les outils et services à valeur ajoutée locaux et distants
- les fournitures du CRDO pour l'interface graphique d'accès aux données
- etc.

7. Acronymes et définitions

7.1. Glossaire

Terme	Définition
Archivage électronique	Voir Archivage numérique
Archivage numérique	Ensemble des actions nécessaires au fonctionnement d'une Archive en charge d'informations sous forme numérique.
archive	Le terme 'archive' sans majuscule désignera un document d'archive.
Archive (ou service d'archives)	Organisation chargée de conserver l'information pour permettre à une Communauté d'utilisateurs cible d'y accéder et de l'utiliser (OAIS)
Authentification	Procédé visant à vérifier l'identification d'une personne physique par des moyens techniques, tels que mot ou phrase de passe, un code secret, une réponse à un défi ou encore une sécurisation numérique (Certificat).
Contenu d'information	Ensemble d'informations constituant l'objet principal de la pérennisation (OAIS)
Donnée	Représentation formalisée de l'information, adaptée à la communication, à l'interprétation ou au traitement. par exemple : une séquence de bits, un tableau de nombres, les caractères d'une page, un enregistrement audio, etc.
Empreinte (empreinte numérique ou condensat ou hash)	Résultat d'une fonction de hachage appliquée sur une chaîne de caractères de longueur quelconque visant à réduire celle-ci en une donnée de longueur fixe représentative de cette chaîne de caractères. L'empreinte est l'un des éléments permettant de vérifier l'intégrité d'un document, d'un flux, d'un lot, d'une transmission... (comparaison d'empreintes).
Entité « Accès » (Access)	Entité de l'OAIS regroupant les fonctions et services de mise à disposition des utilisateurs des collections archivées.
Entité « Administration » (Administration)	Entité de l'OAIS regroupant les fonctions et services requis pour une supervision continue du fonctionnement des autres entités de l'OAIS.
Entité « Entrées » (Ingest)	Entité de l'OAIS regroupant les fonctions et services qui : <ul style="list-style-type: none"> • prennent en charge les Paquets d'informations à verser (SIP) livrés par les Producteurs, • préparent les Paquets d'informations archivés (AIP) en vue de leur stockage, • et assurent la bonne intégration dans l'Archive de ces AIP et de leur Information de description.
Entité « Gestion de données » (Data Management)	Entité de l'OAIS regroupant les fonctions et services dédiés à l'alimentation, la maintenance et la mise à disposition d'informations aussi variées que les catalogues et inventaires de ce qu'il est possible de se procurer auprès de l'Entité « Stockage », les algorithmes de traitement applicables aux données récupérées, les statistiques de consultation, les éléments de facturation, les Demandes d'abonnement, les contrôles de sécurité, les plannings, la politique et les procédures propres à l'OAIS.
Entité « Stockage » (Archival Storage)	Entité de l'OAIS regroupant les fonctions et services utilisés pour le stockage et la récupération des Paquets d'informations archivés (AIP).
Information	Toute connaissance pouvant être échangée. Lors de l'échange, elle est représentée par des données. Exemple : une chaîne de bits (les données) accompagnée d'une description permettant d'interpréter cette chaîne de bits comme des nombres représentant des mesures de températures en degrés Celsius (Information de représentation).
Information de contexte	Information qui décrit les liens entre un Contenu d'information et son environnement. Elle inclut entre autres les raisons de la création de ce Contenu d'information et son rapport avec d'autres Objets-contenu d'information (OAIS)
Information de description	Ensemble d'informations, constitué principalement de Descriptions de paquet, et fourni à l'Entité « Gestion de données » pour aider les Utilisateurs à rechercher,

	commander et récupérer des informations de l'Archive (OAIS)
Information d'empaquetage	Information permettant de relier et identifier les composants d'un Paquet d'informations. Par exemple les informations de volume et de répertoire dans un CD-ROM conforme à la norme ISO 9660, permettant d'accéder aux fichiers supports du Contenu d'information et de l'Information de pérennisation (OAIS).
Information d'identification	Information qui définit, et si nécessaire, décrit le ou les mécanismes d'attribution des identifiants au Contenu d'information. Elle inclut aussi les identifiants qui permettent à un système externe de se référer sans équivoque à un Contenu d'information particulier. Exemple : un ISBN (International Standard Book Number).
Information d'intégrité	Description des mécanismes et des clés d'authentification garantissant que le Contenu d'information n'a pas subi de modification sans que celle-ci ait été tracée. Par exemple, le code CRC (contrôle de redondance cyclique) pour un fichier (OAIS).
Information de pérennisation	Information nécessaire à une bonne conservation du Contenu d'information, et qui peut être décomposée en Informations de provenance, d'identification, d'intégrité et de contexte (Preservation Description Information - PDI dans le Modèle OAIS)
Information de provenance	Information qui documente l'historique du Contenu d'information. Cette information renseigne sur l'origine ou la source du Contenu d'information, sur toute modification intervenue depuis sa création et sur ceux qui en ont eu la responsabilité (OAIS)
Information de représentation	Information qui traduit un Objet-données en des concepts plus explicites. Par exemple, la définition du code ASCII décrit comment une séquence de bits (un Objet-données) est convertie en caractères (OAIS)
Information de structure	Information qui explique la façon dont d'autres informations sont organisées. Elle établit par exemple une correspondance entre les trains de bits et les types de données courants sur ordinateurs (tels que caractères, nombres, pixels ou agrégats de ces types tels que chaînes de caractères et tableaux)
Long terme	Période suffisamment longue pour qu'il soit nécessaire de prendre en compte les changements technologiques, et notamment la gestion des nouveaux supports et formats de données ainsi que l'évolution de la communauté d'utilisateurs. Cette période n'est pas limitée dans le temps.
Objet données	Objet physique ou Objet numérique.
Objet information	Objet-données avec son Information de représentation.
Paquet d'information	Association du Contenu d'information et de son Information de pérennisation, destinée à faciliter la conservation du Contenu d'information. A ce Paquet d'informations est aussi associée une Information d'empaquetage utilisée pour circonscrire et identifier le Contenu d'information et son Information de pérennisation (OAIS).
Paquet d'information archivé (Archival Information Package - AIP)	Paquet d'informations conservé dans une Archive et constitué d'un Contenu d'information et de l'Information de pérennisation (PDI) associée.
Paquet d'informations à verser (Submission Information Package - SIP)	Paquet d'informations livré par le Producteur à l'Archive pour l'élaboration d'un ou plusieurs Paquets d'informations archivés (AIP).
Paquet d'informations diffusé (Dissemination Information Package - DIP)	Paquet d'informations reçu par l'Utilisateur en réponse à sa requête à l'Archive. Ce paquet provient d'un ou de plusieurs Paquets d'informations archivés (AIP).
Producteur	Toute personne ou système client qui fournit des informations à pérenniser. Il peut s'agir d'autres Archives, ou de personnes ou systèmes internes à l'Archive (OAIS). Voir <i>Service versant</i>
Transfert	Transmission effective des SIP entre le Producteur et l'Archive.
Utilisateur	Personne ou système, en relation avec les services de l'Archive pour trouver des informations archivées présentant un intérêt, et pour accéder au détail de ces informations.

7.2. Abréviations

Abréviation	Nom détaillé
AIP	Paquet d'information archivé (Archival Information Package) (terminologie OAIS)
BnF	Bibliothèque nationale de France
CCLE	Cognition Langues Langage Ergonomie (laboratoire)
CCSD	Centre pour la Communication Scientifique Directe
CCSDS	Comité Consultatif pour les Systèmes de Données Spatiales
CINES	Centre Informatique National de l'Enseignement Supérieur
CLAPI	Corpus de LAngue Parlée en Interaction
CNES	Centre National d'Etudes Spatiales
CNRS	Centre National de la Recherche Scientifique
CRDO	Centre de Ressources pour la Description de l'Oral
DAF	Direction des Archives de France
DC	Dublin Core (format standard de métadonnées)
DGLFLF	Délégation Générale à la Langue Française et aux Langues de France
DGRI	Direction générale de la recherche et de l'innovation (ministère de l'enseignement supérieur et de la recherche)
DGES	Direction générale de l'enseignement supérieur (ministère de l'enseignement supérieur et de la recherche)
DIP	Paquet d'information diffusé (Dissemination Information Package) (terminologie OAIS)
DIS	Direction de l'Information Scientifique
ERSS	Équipe de Recherche en Syntaxe et Sémantique
HAL	Hyper Articles en Ligne
ILF	Institut de Linguistique Française
ICAR	Interactions, Corpus, Apprentissages, Représentations
IN2P3	Institut national de Physique Nucléaire et de Physique des Particules
LACITO	Laboratoire de langues et civilisations à tradition orale
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
PAIMAS	Producer archive interface methodology abstract standard (norme ISO 20652). Il s'agit d'une norme qui définit un cadre méthodologique (définitions des grandes étapes de travail) et un ensemble d'actions à prendre en compte dans la relation entre un producteur de document et un service d'archives en charge de conserver ces documents.
PDF	Portable Document Format
PDF/A	Portable Document Format/Archive (norme ISO définie à partir de la version 1.4 de PDF afin de répondre aux besoins de l'archivage long-terme des documents. Un fichier PDF/A contient toutes les polices de caractère dont il a besoin pour restituer le document. Il n'est pas, dans certaines limites, dépendant de l'environnement).
PDI	Information de pérennisation (Preservation Description Information) (terminologie OAIS)
PFC	Phonologie du Français Contemporain
PIN	Pérennisation des Informations Numériques (groupe de travail de l'association Aristote)
SGML	Standard Generalized Markup Language
SIP	Paquet d'Information à verser (Submission Information Package) (terminologie OAIS)
TUL	Typologie et Universaux Linguistiques
UML	Unified Modeling Language
XML	Extensible Markup Language

8. Références bibliographiques

[Barring 08]	Olof Barring, <i>Hosting of IT services and data for Human and Social sciences in France, version 1.0</i> , janvier 2008. Rapport final de l'étude commandée par le TGE Adonis au CERN. Document non public, communicable par le TGE Adonis sur demande.
[OAIS 02]	CCSDS, 650.0-B-1, <i>Reference Model for an Open Archival Information System (OAIS)</i> , ISO 14721, janvier 2002 http://public.ccsds.org/publications/archive/650x0b1.pdf ISO 14721, <i>Reference Model for an Open Archival Information System (OAIS)</i> , 2003, Genève CCSDS, 650.0-B-1(F), <i>Modèle de référence pour un Système ouvert d'archivage d'information (OAIS)</i> , ISO 14721, mars 2005 http://public.ccsds.org/publications/archive/650x0b1(F).pdf
[PAIM 04]	CCSDS, 651.0-B-1, <i>Producer archive interface methodology abstract standard</i> , ISO 20652, mai 2004 http://public.ccsds.org/publications/archive/651x0b1.pdf

9. Annexes

9.1. Volumétrie

Volumétrie actuelle (mars 2008)

Total occupé à ce jour: environ 2 To (= les entrées de 2 ans)

Fichiers audio

- effectif : 3500
- tailles de fichiers (environnement de conservation) : de 1 Mo à 4 Go

Fichiers vidéo

- effectif : 70
- tailles de fichiers (environnement de conservation) : de 8 Go à 20 Go

Fichiers d'annotation

- text/plain : 20
- text/xml : 500
- application/pdf : 100

taille des fichiers négligeable : quelques Ko a quelques centaines de Ko

Quelle estimation des évolutions des volumes et des usages, à 2 ans, à 5 ans ?