

## Annexe à la lettre d'intention CRDO

La création officielle du CRDO nécessite la convergence des deux projets de préfiguration. Dans l'attente de la mise en place du Conseil scientifique du CRDO unifié, les deux laboratoires porteurs des projets de préfiguration constituent un « *Comité CRDO unifié* ». Selon les termes de la lettre du Chargé de mission pour les Sciences du Langage INSHS du 12 avril 2010 :

« *Pour une période transitoire, en attendant que le Conseil Scientifique ait pu se réunir et désigner son président et son secrétariat, la responsabilité des affaires et des initiatives dans le cadre décrit ci-dessus est confiée aux deux laboratoires où le CRDO a été initialement créé : Le LPL d'Aix en Provence et le LACITO de Villejuif. Ces deux laboratoires sont donc chargés d'organiser conjointement la mise en place du Conseil Scientifique, et le suivi des affaires jusqu'au fonctionnement entier de la nouvelle gouvernance.* »

### **Description des actions**

Le CRDO est un service accessible aux laboratoires de recherche et aux chercheurs indépendants souhaitant mutualiser leurs données orales tout en bénéficiant d'un archivage pérenne selon le modèle OAIS (*Open Archival Information System*).

Développé en 2006 à partir d'une initiative conjointe de la Direction de l'Information Scientifique et du département scientifique Homme et Société du CNRS, le CRDO s'est appuyé sur deux propositions portées respectivement par les laboratoires Lactito et LPL : équipes « Gestion documentaire et réservoir de données » et « Ressources et outils d'analyse ».

Les deux projets de préfiguration ont travaillé à la mise en œuvre du modèle OAIS dans le cadre du projet pilote de mutualisation et archivage pérenne de données orales initié par le TGE-Adonis :

[http://www.tge-adonis.fr/wiki/index.php/Accueil\\_Projet\\_pilote](http://www.tge-adonis.fr/wiki/index.php/Accueil_Projet_pilote)  
<http://www.tge-adonis.fr/wiki/uploads/9/9d/ArchivageMutualise-document-synthese-v1.3.pdf>

Ce dispositif fait intervenir le *Centre informatique de l'enseignement supérieur* (CINES) et le *Centre de calcul de l'Institut national de physique nucléaire et de physique des particules* (CC-IN2P3) qui assurent respectivement les fonctions d'archivage pérenne et de diffusion des ressources.

L'historique détaillé des développements de la plateforme de l'équipe aixoise est disponible sur la page : <http://crdo.fr/wiki/Developpement/Journal>. L'historique d'archivage de l'équipe aixoise (en phase de test) peut être consulté ici : [http://crdo.fr/archivage\\_hist.php](http://crdo.fr/archivage_hist.php)

Le site de l'équipe parisienne est <http://crdo.risc.cnrs.fr/>

Cette lettre d'intention concerne le futur du CRDO unifié.

Nos propositions et besoins sont les suivants :

- 1) Préparation de la fusion des projets Lacito et LPL
- 2) Elargissement du champ disciplinaire du CRDO
- 3) Valorisation des objets et des équipes de recherche
- 4) Implication des laboratoires producteurs
- 5) Choix concernant l'authentification des utilisateurs et les licences
- 6) Le CRDO en tant que service intégré à CLARIN
- 7) Elargissement de l'équipe
- 8) Travaux ponctuels
- 9) Travaux récurrents
- 10) Besoins

## 1. Fusion des projets Lacito et LPL

Les méthodes de catalogage et d'archivage des deux préfigurations ont été guidées par la nature des corpus traités : pour simplifier, le type de documents traités par le LACITO/ « préfiguration parisienne » concerne des récits, et la DTD XML a été conçue en fonction de ces documents ; les documents LPL sont des ensembles documentaires en phonétique, aux structures beaucoup plus diverses. Les deux fonds ne peuvent en l'état actuel être fusionnés. Des ateliers LPL-LACITO seront organisés à partir de mai 2010, afin d'envisager de façon très détaillée le futur « CRDO unifié », en abordant de front les questions techniques liées à la gestion des deux ensembles documentaires en constitution dans les deux projets de préfiguration. Les déposants auprès de la préfiguration parisienne du CRDO étant nombreux, la consultation sera élargie à ces partenaires. Une question centrale en vue de la fusion concerne l'indexation des objets labellisés « CRDO » (un changement des identifiants actuels sera sans doute nécessaire ; cela se fera en concertation avec le CINES), ainsi que les pratiques associées à leur diffusion : licences et suivi des utilisateurs.

## 2. Elargissement du champ disciplinaire du CRDO

Le CRDO était en 2006 une initiative des SHS. Les développements effectués par l'équipe aixoise ont pris avantage de la multidisciplinarité des travaux en sciences du langage. Il a été convenu notamment que les données primaires ne se limiteraient pas à des enregistrements sonores de parole mais seraient étendues à tous les signaux associés aux pratiques langagières, qu'il s'agisse de paramètres articulatoires ou de données visuelles permettant l'étude de la mimo-gestualité.

Un autre élargissement a été amorcé à partir des corpus de linguistique de terrain dont la mutualisation permet leur réutilisation par des chercheurs de disciplines extérieures aux sciences du langage : sociologie, anthropologie, histoire, psychologie, musicologie, philosophie etc.

Ces extensions du champ disciplinaire ne sont pas simples à mettre en œuvre. Alors que les formats de données numériques peuvent s'appuyer sur tous les standards validés par la plateforme d'archivage, les métadonnées descriptives sont encore structurées à partir du schéma OLAC (Open Language Archives Community) conçu pour la linguistique textuelle. La question se pose de savoir dans quelle mesure la diversité des données doit être reflétée par des intitulés explicites dans les métadonnées, pour refléter le fait que tel ou tel dépôt concerne des disciplines telles que la phonologie ou la prosodie.

Un travail devra être entrepris avec les producteurs de données orales autres que linguistiques afin que leurs dépôts bénéficient au mieux de la visibilité offerte par le CRDO. **En pratique, la délimitation des priorités est une tâche importante, afin que le CRDO nouvellement fondé ne soit pas submergé par des demandes dépassant ses capacités.**

L'atelier abordera la question du modèle de métadonnées ISOcat (norme ISO 12620), et reprendra la question de l'implémentation de Dublin Core et OLAC dans le futur CRDO ; classification des documents du futur CRDO dans la nomenclature internationale.

## 3. Valorisation des objets et des équipes de recherche

La préfiguration aixoise a mis en place plusieurs dispositifs spécifiques pour améliorer la visibilité des objets archivés et distribués. Ces dispositifs se traduisent par des relations qui permettent d'associer à tout objet :

- un espace wiki « illimité »
- des publications
- des liens vers les pages présentant des programmes de recherche dont ils sont issus
- des liens vers des objets associés : collections, etc.

L'espace wiki est actuellement sous PhpWiki, un logiciel open source PHP/MySQL qui permet l'archivage de l'intégralité du site (après exportation et conversion en UTF8) mais n'est plus entretenu par ses programmeurs. Un script sera utilisé pour migrer l'intégralité du wiki vers PmWiki, un environnement plus performant qui gère de simples fichiers texte sans base de données.

Le site de la préfiguration aixoise intègre une base de données de références bibliographiques associées aux objets déposés. Cette base devra faire l'objet d'une interface plus sophistiquée pour gérer un nombre rapidement croissant d'entrées.

## 4. Implication des laboratoires producteurs

Dans la préfiguration aixoise, chaque objet peut être déposé à titre personnel ou sous la bannière du laboratoire de recherche dont dépendait l'auteur au moment de sa production. Ces laboratoires sont désignés comme « laboratoires producteurs » et répertoriés sur le site : <http://crdo.fr/labs/>

Des propositions seront formulées pour les laboratoires producteurs pour l'indexation des documents. Le futur CRDO n'a pas vocation à prendre en charge le développement d'interfaces de consultation spécifiques : ce sont les systèmes d'information des laboratoires, ou des « communautés de chercheurs » dans un domaine donné (exemple : acquisition de la parole par les enfants, étude de langues rares/peu documentées, données articulatoires...), qui développeront des interfaces dédiées à un certain type de données. (Exemple : site Archivage du LACITO ; projet inter-laboratoires en cours d'élaboration « Langues du monde ».) Mais pour le bon fonctionnement de ces interfaces, les modalités de consultation doivent être préparées en amont. Le futur CRDO devra permettre des requêtes élaborées : streaming d'un extrait, etc. Ce travail devra être entrepris avec un spécialiste de l'environnement *Fedora Commons* utilisé pour la plateforme de diffusion au CC-IN2P3. Des travaux sont en cours à cet effet au LPL.

## 5. Authentification des utilisateurs et licence

La question de l'accès et de l'authentification doit également être abordée. Dans ce domaine, il existe à l'heure actuelle de grandes différences entre les deux préfigurations.

L'état de la réflexion dans la préfiguration aixoise est le suivant : en décembre 2009, le CC-IN2P3 s'est impliqué dans le développement d'un système d'authentification centralisé des utilisateurs compatible avec celui implémenté sur RENATER. Ce système permettra à un utilisateur de s'identifier sur un service (le CRDO en l'occurrence) en saisissant son identifiant et mot de passe dans le même environnement que d'autres services disponibles au CNRS ou dans les universités. Le système utilisé (OpenSSO) permettra en outre de réaliser une authentification via des fédérations d'universités extérieures à RENATER pourvu que leurs systèmes soient compatibles avec le protocole SAML 2. La maquette est décrite sur la page :

[http://www.tge-adonis.fr/wiki/index.php/Etude\\_préalable\\_et\\_maquette](http://www.tge-adonis.fr/wiki/index.php/Etude_préalable_et_maquette)

Le module fonctionne depuis fin mars 2010 et son intégration expérimentale au site aixois est imminente. Elle nécessitera toutefois une réflexion sur le profil minimal d'utilisateur permettant de spécifier les droits d'accès aux diverses catégories d'objets.

Cette question figurera parmi les priorités des ateliers LPL-LACITO.

## 6. Le CRDO en tant que service intégré à un réseau européen

Le CRDO a vocation à fonctionner comme un service ouvert à l'ensemble de la communauté scientifique, offrant aux laboratoires producteurs des facilités pour la mise en forme d'objets (corpus, ressources, outils) en vue de leur archivage pérenne (total ou partiel) au CINES et de leur distribution (totale ou partielle) sur la plateforme développée au CC-IN2P3.

Ce fonctionnement est comparable à celui de HAL. Toutefois, de nombreuses contraintes liées à la diversité et la complexité des objets et des formats de données numériques ne permettent pas en général le traitement automatique des dépôts.

Exemples : (1) un objet peut contenir plusieurs dizaines de fichiers structurés dans une hiérarchie adaptée à son exploitation par les équipes de recherche, ce qui soulève la question de savoir si chaque fichier doit être déposé et référencé indépendamment (le lien étant établi via les métadonnées) ou si l'ensemble doit être déposé comme un dépôt unique (ce qui compliquerait l'action de « pointer » vers un des fichiers en particulier). (2) les noms de fichiers peuvent provenir de systèmes informatiques utilisant

une graphie extra-européenne (codée en Unicode UTF8) ; (3) les droits d'accès et conditions d'utilisation peuvent être différents pour divers types de fichiers.

La préfiguration aixoise a donné la priorité aux développements du service permettant de gérer une très grande diversité d'agencements et de traitements de fichiers et d'objets, afin de faciliter leur réutilisation par des chercheurs de disciplines les plus diverses. La méthode consiste à imposer un minimum de contraintes aux producteurs tout en les guidant vers des solutions optimales compatibles avec l'archivage pérenne (par exemple, le choix d'un taux de compression de fichiers audio/vidéo permettant de préserver la contiguïté des séquences enregistrées.) La préfiguration parisienne a privilégié les fichiers correspondant à la DTD initialement développée au LACITO.

Tous les développements du service doivent donner la priorité à la visibilité des objets sur les sites des laboratoires producteurs, ainsi que la possibilité pour eux de développer des services web effectuant des traitements sur ces objets, comme proposé au § 3. Le CRDO doit donc servir de simple relais (« service versant ») entre les laboratoires producteurs (ou, occasionnellement, un producteur indépendant) et les plateformes d'archivage et de distribution.

S'il est envisageable que d'autres services versants deviennent opérationnels pour le dépôt de données orales, il faudrait que tous ces services respectent une indexation unique des objets. Cette unification permettrait de consulter un *repository* pour retrouver le service versant et/ou le laboratoire producteur d'un objet puis lancer une requête standardisée sur cet objet comme précisé au § 4.

Le 23 janvier 2010, la préfiguration aixoise du CRDO est inscrit dans le réseau européen CLARIN. A cette occasion, le groupe aixois s'est inscrit à plusieurs groupes de travail : WG 2.1, 2.2, 2.3, 2.4 (voir <<http://www.clarin.eu/og/all?filter0=2.>>). Cette participation vise à intégrer le CRDO comme une « tête de pont » de l'archivage pérenne des données orales en France, mais elle nécessitera de dégager assez de temps de travail pour devenir efficace. Parmi les priorités des ateliers LPL-LACITO figurera une réflexion sur la façon dont l'intégration en cours de la préfiguration aixoise dans CLARIN peut être transposée au « CRDO unifié », en fonction des choix techniques qui y seront effectués. Il est souhaitable que le « CRDO unifié » bénéficie des acquis des deux préfigurations, dans toute la mesure du possible ; son intégration à CLARIN revêt une grande importance pour son développement.

## 7. Elargissement de l'équipe

Les modalités d'élargissement vont dépendre des décisions du Conseil Scientifique à réunir, mais on peut indiquer d'emblée deux champs nécessaires.

Il est souhaitable de mettre en place une équipe technique, fonctionnant à distance, dont certains membres auront des droits d'administrateurs sur certaines fonctions du CRDO. D'autres membres pourraient effectuer des tâches nécessitant une expertise particulière : conversion de formats, interactions avec les producteurs et responsables de projets en vue d'un conditionnement optimal des objets pour l'archivage pérenne et la mutualisation, recommandations sur les droits d'accès, etc.

S'il devient nécessaire d'établir des priorités d'archivage en fonction des quotas de volumes de données accordés par la plateforme du CINES, le Comité Scientifique devra soit assurer lui-même la responsabilité des choix en fonction de critères précis, soit déléguer la tâche à un comité, en se réservant le contrôle et la responsabilité.

## 8. Travaux ponctuels

- Mise en place de l'authentification centralisée
- Développement de services en direction des laboratoires producteurs
- Amélioration des métadonnées descriptives
- Etude des problèmes spécifiques du dépôt de données orales de domaines extérieurs aux sciences du langage
- Meilleure gestion de la base de références bibliographiques
- Migration du wiki de la préfiguration aixoise vers le logiciel PmWiki
- Mise en place d'un site CRDO. Les sites actuellement en ligne (avril 2010) sur [crdo.fr](http://crdo.fr) et [crdo.risc.cnrs.fr](http://crdo.risc.cnrs.fr) constituent des préfigurations. Ni l'une ni l'autre de ces préfigurations ne présente les caractéristiques requises pour être promue en l'état au statut de « CRDO unifié », du fait de

différences techniques qu'aborderont les ateliers LPL/LACITO (organisés à partir de mai 2010) .  
Le site du CRDO unifié sera mis en place sur la base de ces ateliers.

## 9. Travaux récurrents

- Mise en forme des objets en collaboration avec leurs producteurs
- Conversion/recyclage de fichiers refusés par le validateur du CINES
- Planning d'archivage en fonction des priorités
- Participation aux groupes de travail de CLARIN

## 10. Besoins

Nous avons besoin pour assurer un passage à l'échelle et la mise en place d'une production stabilisée d'un soutien en termes de fonctionnement. Ce soutien porte tout d'abord sur le développement informatique du portail et des services associés. Nous sollicitons pour cela un assistant ingénieur informaticien. Par ailleurs, le passage en mode de production nécessitera dans la période de démarrage un suivi particulier des laboratoires producteurs à la fois pour les convaincre du service puis pour les accompagner dans leur dépôt. Là encore, un assistant ingénieur sera chargé de ce travail.

Fonctionnement	Assistant ingénieur chargé du développement de la plateforme	33.000 €
	Assistant ingénieur chargé du suivi des ressources : contact avec les laboratoires producteurs, formatage des ressources, mise en place des relations producteurs/crdo, etc.	33.000 €
Equipement	2 postes de travail	2.000 €
	Baie de stockage	3.000 €
	Consommables	1.000 €
Total		72.000 €

Ce financement est demandé par les deux laboratoires au nom du Comité scientifique qui se mettra en place dans les prochains mois ; il permettra au projet de commencer effectivement à fonctionner dans l'attente de cette mise en place. L'affectation de l'un des deux CDD à chacun des deux laboratoires est envisagée.