

Les travaux de la composante parisienne du crdo: « Gestion documentaire et réservoir de données »

Ce document fait un rapide bilan des actions menées jusqu'à présent dans notre centre de ressources. Notre groupe s'active fortement dans la perspective de mise à disposition de corpus existants. Un important travail a déjà été effectué pour la mise en place d'une architecture de stockage et diffusion de ces corpus qui respectent au plus près les standards internationaux. Cette architecture est déjà opérationnelle et nous permet déjà d'accueillir facilement les nouveaux corpus de toute la communauté.

Contrôle et acceptation des ressources ;

- Organisation du dépôt en-ligne des ressources:
 1. Nature des ressources: Les ressources susceptibles d'être soumises sont des enregistrements oraux (audio, vidéo ou mesures physiologiques) accompagnées éventuellement de leurs annotations directes. Les contraintes de codage et de formatage sont expliquées sur le site web. Une aide peut éventuellement être apportée pour normaliser des données qui ne le sont pas.
 2. Gestion des comptes utilisateurs et des droits associés: Pour pouvoir déposer une ressource dans l'archive, il faut s'enregistrer au préalable afin que l'on sache qui est le déposant (un chercheur d'une institution de type académique: CNRS, Université, etc.)
 3. Deux types de dépôt sont possibles: soit par un formulaire en-ligne, soit en suivant des instructions pour décrire en xml les ressources.
 4. Contrôle qualité: Une fois la ressource soumise, celle-ci est examinée par une personne du centre pour une vérification de forme (format et codage du fichier). La description des ressources est, elle aussi, examinée par un documentaliste du centre qui va en vérifier la complétude et l'exactitude. Si un manque ou une erreur est détecté tant dans la description de la ressource que dans le codage ou le formatage de celle-ci, une personne du centre prendra contact avec le déposant via son email.

Stockage, maintenance et sauvegarde des ressources ;

Du matériel de stockage a été acheté par le centre afin de pouvoir conserver et diffuser les ressources de la communauté. Nous distinguons 2 environnements:

1. Un environnement de conservation qui stocke les fichiers dans un format normalisé et d'une qualité optimale. Les recommandations de IASA sont pour l'audio par exemple (96KHz, 24bits)
 1. Un environnement de diffusion qui stocke et diffuse des copies dégradées des ressources (22KHz, 16bits, mono)
- Le matériel de stockage (une baie de 7 To effectif avec une redondance par un système RAID) est hébergé dans l'unité de service RISC qui a accepté d'en faire l'administration. D'autres disques ont été achetés et placés dans des serveurs de Villejuif afin d'avoir à tout moment plusieurs lieux physiques de stockage l'ensemble des données.
 - Nous avons aussi acheté un serveur dédié à la gestion de l'archive et sa diffusion. Ce serveur est lui aussi hébergé au RISC qui en assure la maintenance et l'administration système.
 - Un site web a été créé (mise en activité février 2006) pour présenter la composante parisienne du crdo et servir de portail sur les ressources (<http://crdo.vjf.cnrs.fr>). Ce site présente l'organisation, les missions et les projets en cours de cette composante. Il offre aussi une interface d'accès aux ressources.

Productions et mises à disposition :

Un réservoir de ressources a été créé (en février 2006) basé entièrement sur les technologies W3C. Un programme implémentant le protocole OAI-PMH a été développé afin de disséminer les métadonnées. Il a été enregistré auprès des organisations OAI et OLAC comme « data provider ». Ce réservoir est

« moissonné¹ » régulièrement notamment par les moteurs de recherche Google et OAIster. Un moteur de recherche spécifique a été développé, ainsi qu'une interface de consultation des ressources. Ce réservoir contient à ce jour plusieurs centaines d'heures d'enregistrements dont plusieurs dizaines d'heures en libre accès. Afin de couvrir l'ensemble des besoins exprimés par les chercheurs, nous avons distingué 3 niveaux de protection : 1) données et métadonnées en libre accès, 2) métadonnées publiques mais données accessibles par authentification, 3) données et métadonnées accessibles par authentification.

Les ressources déposées proviennent de différentes institutions du monde académique: les Universités d'Orléans², de Nantes³, et de Paris 3⁴, ainsi que les laboratoires CNRS: CELIA⁵, CRLAO⁶, LCF⁷, ERSS⁸, LMS⁹, ICAR et l'Institut Gardette¹⁰ et le LACITO¹¹.

De nombreuses heures d'enregistrement sont aussi en cours de numérisation ou de dépôt, en provenance de ces mêmes institutions ainsi que d'autres telles que l'EHESS (enregistrements de langues de Nouvelle-Calédonie), un ancien atlas linguistique (langues d'Oïl) fait par des enseignants et des chercheurs de diverses appartenances, des corpus de langage enfantin, etc.

Plusieurs travaux de valorisation de corpus ont été entrepris.

- Pour le corpus ESLO, une maquette de consultation a été développée mixant dans les requêtes des critères socio-linguistiques et des critères portant sur la transcription.
- Un site portail web pour le projet DALLITH (contrat PICS passé entre le CNRS/Lacito et l'Université de Virginie USA) pour donner un accès à des ressources venant des deux partenaires et concernant les langues de la zone himalayenne.
- Le site du Lacito a été revu pour s'interfacer avec la nouvelle architecture mise en place par le CRDO

Un travail de standardisation est en cours avec le projet PFC afin de lui permettre de diffuser ses métadonnées avec le protocole OAI.

Un autre travail de standardisation sera mené par le CRDO dans le cadre du projet VARILING qui vient d'obtenir un financement ANR. Il s'agit de la continuation d'un groupe de travail CatCod « Catalogage et Codage pour les corpus oraux ». Les objectifs principaux sont: a) de rassembler la communauté autour de ces problèmes afin que chacun connaisse les pratiques existantes en la matière, b) de trouver un noyau commun qui peut faire l'objet d'un consensus, d) de formaliser ce noyau pour le proposer à un organisme de normalisation (TEI dans le cadre d'une refonte du chapitre sur la transcription de l'oral) afin de stabiliser ces acquis.

Les projets en cours :

- Un ensemble d'extraits de corpus oraux numérisés dans le cadre du contrat entre la DGLFLF et les deux fédérations de linguistiques TUL et ILF pour un portail sur le « français et les langues de France » pour le ministère de la Culture. LA DGLFLF poursuit le financement de ce programme sur 2007

1 Terme OAI qui

2 Laboratoire ICAR: Langues: angolaise, sãotomense et français (notamment le corpus ESLO (350 heures d'une enquête en socio-linguistique de la fin des années 60)

3 Laboratoire de Linguistique de Nantes: Du kabyle

4 Principalement du français mais aussi de l'allemand et du vietnamien. Certains enregistrements sont des électro-glotto-graphes.

5 Des langues de Guyane, du Surinam et de Bolivie

6 Du japhug (langue de Chine)

7 Du créole réunionnais

8 Du français extrait du grand corpus PFC (Phonologie du français contemporain) pour une diffusion sur le portail de la DGLFLF sur le « français et les langues de France »

9 Un corpus de judéo-espagnol parlé en France d'environ 70 heures

10 Une vidéo et un enregistrement sonore extraits de la base CLAPI pour une diffusion sur le portail de la DGLFLF sur le « français et les langues de France »

11 De nombreuses langues de Nouvelle-Calédonie, du Népal, du Caucase, d'Afrique, etc.

- L'enquête Socio-Linguistique d'Orléans (ESLO) d'un volume de 350 heures d'enregistrement de français de la fin des années 60 (Université d'Orléans)
- Un corpus sur l'apprentissage du langage par des enfants. Etude longitudinale sur plusieurs enfant de 0 à 3 ans avec des enregistrements vidéo d'une heure tous les mois. (Coordination MoDyCo - ENS-LSH)
- Corpus sur des langues Afroasiatiques (Projet CORPAFROAS avec un financement ANR accepté)
- Un Corpus de Français d'Abijan (Université de Besançon)
- Plusieurs corpus de langues d'Afrique (LLACAN)
- Corpus tibétain d'une centaine de documents (contrat PICS avec l'Université de Virginie).

Un comité de pilotage c'est réunit une première fois avec les représentants de tous les laboratoires de la composante parisienne du crdo + un représentant de la composante LPL-Aix + 2 experts extérieurs (ICAR et CORAL).

Contact Européen :

Une mission a été organisée en mars 2006 pour visiter l'équipe technique de Peter Wittenburg au PMI de Nimègue. Cette visite a débouché sur la volonté de part et d'autre de trouver des terrains de coopération. Un premier terrain pourrait être pour nous de délivrer les métadonnées dans le schémas IMDI, pour eux de moissonner les archives avec le protocole OAI. Nous avons aussi signé avec eux la charte sur les « live archives »¹²

Formations :

Participation à l'école thématique CNRS « Linguistique de Corpus Oraux » de Nantes (19 au 24 juin 2006). De nombreux contacts ont pu être pris à cette occasion avec l'ensemble de la communauté. Ces contacts devraient déboucher rapidement sur plusieurs projets de dépôt et de valorisation de corpus. A la demande des organisateurs, le site de cette école est hébergée sur celui du CRDO (<http://crdo.vjf.cnrs.fr/ecoles/elco>)

Michel Jacobson est intervenu ponctuellement dans différents laboratoires et séminaires afin de présenter l'organisation et les travaux du Centre.

12 http://www.mpi.nl/DAM-LR/flyers/DLRA_Flyer_2006-04-23.pdf