LE CENTRE DE RESSOURCES POUR LA DESCRIPTION DE L'ORAL (CRDO)

http://www.CRDO.fr

Bernard Bel & Philippe Blache

Les centres de ressources numériques (CRN) sont une initiative conjointe de la Direction de l'Information Scientifique et du département scientifique Homme et Société du Centre National de la Recherche Scientifique (CNRS). Le Centre de Ressources sur la Description de l'Oral (CRDO) est le CRN centré sur les ressources orales. En 2006, le CNRS a confié la tâche au LPL et au LACITO de créer le CRDO. Son objectif est de répertorier, centraliser et distribuer les corpus, les ressources et les outils de l'oral, avec une priorité pour le français.

Les travaux du CRDO ont été répartis entre deux groupes de développement travaillant sur des aspects complémentaires du projet :

- le groupe « Ressources et outils d'analyse » du CRDO coordonné par le Laboratoire Parole et Langage (LPL) ;
- le groupe « Gestion documentaire et réservoir de données »[1] coordonné par le Laboratoire des langues et civilisations à tradition orale (LACITO).

Cet article présente les travaux en cours dans le groupe « Ressources et outils d'analyse ».

1. Spécificités du groupe

Il se propose de compléter la mise à disposition de données par les ressources et outils facilitant leur exploitation. Il s'agit de fournir un ensemble d'instruments allant du signal acoustique brut à l'édition et au traitement de ce signal. Un tel service doit donner accès aux informations elles-mêmes, à des outils et méthodes permettant d'en effectuer l'analyse, et autant que possible à des données enrichies par ces outils.

Les dépôts appartiennent pour cela à trois catégories distinctes :

- les « données » qui peuvent être des documents sonores/vidéo bruts ou des séquences d'informations le cas échéant alignées sur ces documents : annotations, transcriptions, mesures de paramètres articulatoires *etc.* ;

1. http://crdo.vjf.cnrs.fr.

- les « ressources » qui sont des ensembles d'informations structurées utilisées dans l'analyse ou la description de la parole : lexiques, tables de fréquences, bases de connaissances *etc.* ;
- les « outils » qui sont des logiciels ou des matériels permettant l'analyse et l'enrichissement des données.

En 2006-2007, les travaux de ce groupe se sont orientés autour de cinq priorités :

- la mise en place d'une architecture informatique robuste, voir « Serveurs » ;
- la gestion de métadonnées décrivant chaque dépôt aussi finement que souhaité par le déposant, tout en permettant l'exportation des descripteurs linguistiques conformes aux standards internationaux ;
- la valorisation du centre de ressources par le suivi des téléchargements et des ressources ainsi distribuées, voir « Valorisation » ;
- la possibilité de gestion coopérative des métadonnées et des informations (tutoriels *etc.*) associées à chaque dépôt, voir « Groupes » ;
- la mise en place d'un réseau de déposants, utilisateurs et laboratoires impliqués à divers titres dans le projet CRDO.

2. Serveurs

La mise en place du serveur s'est faite en deux étapes avec un recouvrement assurant la continuité du service. Dans un premier temps, une base de données relationnelle a été mise en service pour le dépôt et la gestion directe de métadonnées. Cette base fournit les liens vers les ressources physiques proprement dites qui seront installées sur le serveur.

La structure de la base a évolué avec le double objectif de n'éliminer aucune information a priori et de fournir en sortie les données structurées conformément aux standards *(Dublin Core* et OLAC).

Les données sur les institutions productrices des ressources sont stockées dans une base indépendante (617 institutions)[2] utilisée par d'autres services indépendants du CRDO. La réutilisation de la même base garantit une meilleure mise à jour en impliquant d'autres acteurs. Cette base contient aussi, le cas échéant, l'index de l'institution dans le répertoire du CCSD (serveur HAL)[3].

2. http://teck.lpl.univ-aix.fr/institution/institution-recherche.htm

3. http://import.ccsd.cnrs.fr/doc/?consultLabs

Par la suite, nous avons installé une baie de stockage XRAID (RAID 5) connectée par fibre optique à un serveur Xserve G5 assurant toutes les fonctions dans l'environnement Apache/PHP/MySQL. Ce serveur bénéficie d'une sauvegarde incrémentale sur un site éloigné (sécurité « incendie »).

Un nouveau site entièrement construit à partir d'outils W3C est en fin de développement sur le serveur définitif . L'interface est trilingue : français, anglais et chinois[4].

Le site est doté d'un système d'identification, de sorte que certaines actions sont réservées aux membres inscrits. Tout visiteur pourra visualiser en libre accès la liste des données, outils et ressources répertoriés sur le serveur du CRDO, et afficher la partie publique de leurs métadonnées. Pour les corpus audio et vidéo, des échantillons seront accessibles en lecture. Une prévisualisation des enrichissements des données correspondant à ces échantillons (annotation, analyse prosodique, données articulatoires *etc.*) est aussi prévue. Le libre accès permettra le référencement automatique sur les moteurs de recherche.

Un système de requête permettra de sélectionner des dépôts selon des critères appliqués aux champs de métadonnées.

La nouvelle base de données comprend la trace détaillée des téléchargements (pour usage interne) et un répertoire des publications/valorisations associées à chaque téléchargement (voir « Valorisation »).

Les groupes de travail auront accès à un gestionnaire de contenu *(Content management System,* sous Wiki) pour la rédaction coopérative de pages servant à la description des ressources, à l'élaboration de tutoriaux, ainsi que l'archivage de commentaires d'utilisateurs avisés.[5]

3. Valorisation

La signature de la licence CRDO (voir « Licences ») engage l'utilisateur à signaler au CRDO toute publication ou utilisation *(i.e.* « valorisation ») associée au téléchargement d'une ressource. La nouvelle base de données construit à cet effet un répertoire de ces valorisations en relation avec chaque téléchargement.

La saisie sera faite par l'utilisateur (auteur du téléchargement) à partir de son espace personnel. Le responsable du dépôt sera averti par courrier de toute saisie d'une nouvelle valorisation.

Les valorisations seront visibles dans l'espace public du site.

---

4. Le trilinguisme permet de tester toutes les fonctionnalités avec un codage de texte et d'environnement informatique dans les langues extra-européennes.

5. Certaines pages d'intérêt général continueront à être gérées sur Wikipedia, par exemple : <http://en.wikipedia.org/wiki/Prosody_(linguistics)> et <http://en.wikipedia.org/wiki/INTSINT>

4. Déposants, utilisateurs, droits

Un formulaire d'inscription permet à tout visiteur ou déposant d'être répertorié dans la base de données. Toute inscription validée accorde des droits nouveaux en fonction du statut de la personne inscrite. Ce statut est déterminé par les administrateurs du CRDO après examen et vérification éventuelle des informations fournies à l'inscription : identité, affiliation institutionnelle *etc.*

Plusieurs classes d'utilisateurs ont été définies en fonction de leur statut universitaire et de leurs activités. Le responsable de chaque dépôt définit les droits de téléchargement accordés à ces classes et la/les licences associées à l'utilisation du dépôt (voir « Licences »).

Un espace privé est réservé à chaque déposant/utilisateur pour visualiser/modifier ses informations personnelles, les métadonnées des dépôts effectués à son nom, et les listes de ses publications associées aux téléchargements de données du CRDO.

5. Groupes

Il est possible d'associer à chaque dépôt un groupe de travail. Ce groupe est uniquement caractérisé par un identificateur et un mot de passe, indépendamment de l'identification individuelle. Il sera toutefois nécessaire de s'identifier (en tant que déposant ou/et utilisateur) pour accéder au groupe.

Toute fiche de métadonnées appartenant à un groupe peut être modifiée, soit par le déposant, soit par une personne identifiée ayant accès au groupe.

6. Licences

Avant d'ouvrir le serveur au dépôt ou au téléchargement nous avons estimé nécessaire de définir les conditions d'utilisation des ressources distribuées par ce dispositif. Pour cela nous mettons en place une licence comportant deux parties, dont la première (licence « CRDO ») sera obligatoire et la seconde optionnelle.

Pour ce qui concerne le CRDO, il nous est paru prioritaire d'obtenir (et de mettre à disposition de la communauté scientifique) la trace des utilisations des ressources téléchargées. En signant la licence CRDO, tout utilisateur prendra donc un double engagement :

- signaler clairement l'origine (identificateur unique) dans toute publication ou utilisation (même non commerciale) de la ressource ;
- informer le CRDO (par le biais d'une saisie dans l'espace de métadonnées) de cette publication ou utilisation.

La seconde partie de la licence, qui sera laissée au choix du déposant, sera un formulaire de licence *Creative Commons* (préférable pour les données), LGPLLR *(Lesser General Public License for Linguistic*

Resources, préférable pour les ressources linguistiques) ou encore une licence personnalisée par le déposant.

7. État des données archivées

La phase actuelle d'expérimentation s'est appuyée sur l'intégration à la base de ressources internes au LPL :

- 17 fiches ont été entrées et validées pour effectuer la mise au point, dont 8 corpus, 8 outils et 1 ressource.
- 350 fiches en cours de validation décrivent les corpus disponibles au LPL dont la numérisation et la description font l'objet d'un travail séparé.

Les métadonnées de corpus du français collectées par M. Paul Cappeau (FORELL) dans le cadre du groupe de travail ont été fournies le 13 novembre. Il s'agit d'un répertoire d'environ 150 corpus disponibles dans les institutions de plusieurs pays. Ces métadonnées seront prochainement intégrées à la base de données.

8. Coordination du groupe

- Une réunion du Comité de pilotage du CRDO (32 participants) a eu lieu le 12 septembre 2006 à Lyon (Ecole normale supérieure Lettres et Sciences humaines). Elle a débuté par une présentation des objectifs et des outils du CRDO, ainsi que des spécificités et complémentarités de ses deux groupes de développement. Les programmes du CNRTL et de la DGLFLF ont été exposés par MM. Jean-Marie Pierrel et Olivier Baude. La suite de la réunion a consisté à faire l'état des ressources et outils développés et/ou utilisés par les laboratoires BCL, CORAL, DDL, FORELL, ICAR, ICP, LACITO, LIA, LIMSI, LORIA, LPL, LPP et MODYCO.
- Une liste de discussion a été créée pour le partage d'informations au sein du groupe de travail : <http://mailup.univ-mrs.fr/wws/info/CRDO-liste/>.

9. Les perspectives pour 2007

- La bascule vers le serveur définitif dès que toutes les procédures de création/modification de fiches, validation, communication avec les déposants et gestion du répertoire de déposants et de dépôts seront totalement opérationnelles.
- L'exportation automatique des métadonnées de nature linguistique vers un fichier XML respectant le format OLAC afin que les mises à jour de la base du CRDO soient automatiquement référencées.

- La mise en service des espaces Wiki pour la rédaction coopérative de pages web complétant les métadonnées.
- La mise en place de procédures de consultation rapide avec préaffichage d'extraits des corpus sonores et de leurs enrichissements.
- La mise en place de procédures de gestion des données et des utilisateurs permettant le suivi et la maintenance d'un grand nombre de dépôts.
- L'analyse détaillée des métadonnées afin de compléter la description des corpus avec l'aide des déposants et d'étudier au cas par cas leur déplacement vers le serveur du CRDO.