Ortolang
Open Resources and Tools for LANGuage
Outils et Ressources pour un Traitement Optimisé de la LANGue

Numéro 1, page 1 – Mai 2014

Édito

Du manque émerge le besoin et des moyens naissent les opportunités. Ainsi, l'appel à projets dans le cadre des investissements d'avenir a-t-il permis à ORTOLANG de passer du concept à la réalité.

Voici maintenant près de deux ans que nos équipes unissent leurs compétences et leurs efforts autour d'un objectif commun : créer une infrastructure nationale en réseau offrant un réservoir unique de ressources sur la langue française (corpus, lexiques, dictionnaires, etc.) et d'outils numériques pour son traitement.

Au-delà du respect du calendrier, une autre priorité est inscrite sur la feuille de route d'ORTOLANG : le partage de notre production avec la communauté scientifique et la société civile, et son transfert vers le monde de l'entreprise, en particulier les PMI et les PME.

Ce sont bien ces enjeux qui définissent l'ampleur de notre tâche. Nous pouvons compter sur la complémentarité de nos savoir-faire, la richesse de nos interactions et la force de notre motivation pour relever les défis qui jalonnent notre parcours, et honorer notre engagement. A l'arrivée nous est promise la satisfaction d'une mission accomplie et d'une réussite collective.

Merci aux participants au projet, aux membres des comités scientifique et technique qui nous accompagnent dans cette belle expédition humaine et scientifique.

Jean-Marie PIERREL
Directeur
Équipex ORTOLANG

La langue française et les langues de France compteront avec ORTOLANG

Il existe déjà un certain nombre d'outils et de ressources pour l'étude et la connaissance de la langue française, mais les difficultés actuelles pour repérer et accéder à ces ressources (corpus, dictionnaires, lexiques et outils de traitement) sur notre langue résident tout à la fois dans leur grande dispersion (il n'est pas aisé de savoir quelles ressources sont disponibles et à quels endroits elles sont accessibles) et leur forte disparité, en particulier en termes de codage.

En 2012, l'appel dans le cadre des Programmes d'investissements d'avenir, a permis d'ouvrir de nouvelles perspectives et de valider un projet ambitieux associant de multiples partenaires* autour de la réalisation d'un portail unifié : ORTOLANG (Outils et Ressources pour un traitement Optimisé de la Langue, en anglais : Open Resources and Tools for Language).

L'objectif d'ORTOLANG est pluriel comme le rappelle Jean-Marie Pierrel qui en assure la direction scientifique : *« Il s'agit de hisser au meilleur niveau international la recherche, l'analyse, la modélisation et le traitement automatique de notre langue en s'appuyant sur une véritable mutualisation des compétences et des résultats. La communauté scientifique sera bénéficiaire des ressources et des outils ainsi élaborés, mais l'objectif est également d'une part, de faciliter leur transfert et leur usage vers des partenaires industriels, en particulier des PMI et des PME, et d'autre part, de valoriser le français et les langues de France par le partage de ces connaissances issues de nos laboratoires publics ».*

ORTOLANG, dont on mesure bien les enjeux scientifiques, économiques et sociétaux, est accompagné financièrement par une enveloppe de 2,6 M€ HT. Il s'appuie sur un comité technique représentatif des partenaires du projet qui se réunit toutes les six semaines, un conseil scientifique composé pour deux tiers de membres français extérieurs au projet et pour un tiers d'experts étrangers qui se réunit une fois par an et un comité d'orientation associant les représentants des tutelles des unités impliquées.

La version 0 du portail est maintenant ouverte (www.ortolang.fr). L'architecture matérielle de l'équipement est aujourd'hui opérationnelle et son architecture logicielle poursuivra son évolution jusqu'en en 2016, année où est programmée la phase de plein fonctionnement.

« À terme, l'objectif est aussi de faire d'ORTOLANG, un nœud du réseau CLARIN (Common Language Resources and Technologies Infrastructure : www.clarin.eu) » conclut Jean-Marie Pierrel.

William del-Mancino
Responsable de communication

* Analyse et Traitement Informatique de la Langue Française - Laboratoire Parole et Langage - Modèles, Dynamiques, Corpus - Laboratoire Ligérien de Linguistique - Laboratoire LOrrain de Recherche en Informatique et ses Applications - INstitut de l'Information Scientifique et Technique

Un partenariat entre trois pôles géographiques

Nancy

atiff — Analyse et Traitement Informatique de la Langue Française

Loria — Laboratoire lorrain de recherche en informatique et ses applications

INIST

Aix en Provence

UMR 7309 • CNRS • AMU
PAROLE ET LANGAGE

Paris-Nanterre et Orléans

modyco

Laboratoire Ligérien de Linguistique

Et deux centres de ressources

CNRTL
(Centre National de Ressources Textuelles et Lexicales)

SLDR
(Speech & Language Data Repository)

Les principales caractéristiques d'ORTOLANG

ORTOLANG regroupe des compétences complémentaires en
- sciences du langage (ATILF, LPL, MoDyCo et LLL),
- informatique (LORIA et INIST, mais aussi en partie ATILF et LPL),
- base de données et accès à de l'information scientifique (INIST), et à des ressources linguistiques, à travers les centres de ressources (CNRTL, SLDR).

ORTOLANG s'appuie sur une expérience acquise et sur une bonne insertion tant nationale qu'internationale :
- acquis des centres de ressources et laboratoires partenaires qui alimentent la version initiale de la plateforme avec des ressources et des outils déjà disponibles et dont les compétences recouvrent les trois principaux aspects visés : l'oral, l'écrit et la patrimonialisation des langues de France ;
- implication et cohérence avec la TGIR Huma-Num et avec l'infrastructure européenne CLARIN ;
- cohérence avec les efforts de la DGLFLF et de la BNF sur les aspects patrimonialisation des langues de France.

ORTOLANG est une infrastructure de mutualisation pour la gestion, la pérennisation et la diffusion de corpus et d'outils sur la langue, ces derniers restant bien entendu propriété des déposants.

Les objectifs et les missions

Identification et préparation des données

Une des difficultés actuelles pour repérer des ressources sur notre langue (corpus, dictionnaires, lexiques et outils de traitement) et pour y accéder réside dans leur grande dispersion et leur forte disparité, en particulier en terme de codage. Les premiers objectifs concernent donc :
- le catalogage des ressources à travers des métadonnées normalisées,
- le contrôle et la validation des ressources et des outils,
- l'accompagnement des auteurs de ressources sur les standards, les normes et les recommandations internationales actuelles,
- l'enrichissement de ressources et d'outils.

Pérennisation des ressources

Afin d'assurer la pérennisation des ressources, nous mettons en œuvre trois types d'actions :
- curation des ressources et des outils ;
- stockage sécurisé et maintenance des ressources ;
- archivage pérenne, à travers la solution mise en place par la TGIR Huma-Num en lien avec le CINES.

Diffusion

Un accès permanent 24h/24 et 7j/7 aux données sera assuré, tout en respectant les contraintes de droit d'accès aux données.
Nous prévoyons une aide et un accompagnement des utilisateurs pour déposer sur la plateforme et/ou exploiter ces ressources et outils mutualisés en nous appuyant sur l'expérience des équipes porteuses de l'Equipex et des centres de ressources CNRTL et SLDR appelés à se fondre au sein d'ORTOLANG.

Coopérations

Huma-Num

- ORTOLANG opérateur d'Huma-Num pour le domaine linguistique
- Archivage pérenne assuré par Huma-Num
- Appels à projets communs pour la standardisation de Corpus avec les consortiums *Corpus Ecrit* et *IRCOM*

ILF
INSTITUT DE LINGUISTIQUE FRANÇAISE

ORTOLANG, focalisé sur la Langue Française,

- est complémentaire de la fédération ILF
- sera la plateforme d'accueil pour le projet « corpus de référence du français »

Pour accéder ou suivre ORTOLANG :

www.ortolang.fr

contact@ortolang.fr

Directeur de la publication
Jean-Marie Pierrel

Rédacteur en chef
William del-Mancino

L'architecture matérielle

Implantée à l'INIST, elle repose sur des moyens spécifiquement acquis par ORTOLANG (serveurs, système d'exploitation, disques durs, robotique de sauvegarde) et des moyens INIST partagés (réseau, pare-feu, système de stockage et de sauvegarde (SAN), salles machines). Elle d'appuie sur :

- trois *serveurs* DELL R620 bi-processeurs, disposant chacun de 128 Go de mémoire vive, sur lesquels est implanté un système de virtualisation VMware VSphere Entreprise et le système SUSE Linux Enterprise ;
- *un système de stockage* de disques, acquis par ORTOLANG, d'une capacité totale brute de 50 To insérés dans le sous-système de stockage INIST ;
- *un sous-système de* sauvegarde, acquis par ORTOLANG, avec une robotique dédiée (2 lecteurs LTO6 et une librairie pouvant contenir cinquante cartouches). Ces équipements s'intègrent dans l'infrastructure de stockage et de sauvegarde de l'INIST (SAN) et bénéficient du serveur pilotant les sauvegardes (DELL R910) et de la licence *site* HP Data Protecteur relative à la capacité sauvegardée et aux clients de sauvegarde. La capacité de chaque cartouche étant de 2,5 To (6,2 To pour un facteur de compression de 2.5 apportée par les lecteurs LTO6), la capacité totale de sauvegarde est donc de 125 To (312,5 To compressés).

L'architecture logicielle

L'architecture logicielle d'ORTOLANG s'appuie sur un *centre de diffusion* et des *centres thématiques* directement accessibles par les utilisateurs afin de permettre la navigation dans les collections de ressources ou l'obtention de ressources via des requêtes sur les données ou les métadonnées.

Le centre de diffusion

Couche basse de l'architecture logicielle d'ORTOLANG, il supportera des contraintes de qualité de service (disponibilité maximale) et de gestion des documents permettant d'obtenir le DSA (Data Seal of Approval). Ce centre, entrepôt OAI-PMH sera un dépôt fiable des données assurant les fonctionnalités d'identification pérenne de chaque ressource (Handle), de preuve d'intégrité de la donnée associée à un identifiant, de gestion de versions, d'authentification des utilisateurs (à travers un mécanisme de signature unique, Single Sign On) lors de l'accès à des données à accès restreint.

Les centres thématiques

Partie directement visible pour les utilisateurs, trois *centres thématiques* sont proposés, orientés vers les aspects textuels, oraux et patrimoniaux.
L'ensemble des données hébergées sera visible à partir de chaque centre thématique, mais chacun proposera des méthodes de navigation et des interfaces de recherche et de visualisation qui leur sont spécifiques.
Les centres thématiques sont également les interlocuteurs des déposants. Il est de leur responsabilité de mettre en forme données et métadonnées avant transmission au centre de diffusion. Les *centres thématiques* doivent aussi permettre aux chercheurs de se constituer des corpus de travail de façon transparente. Ils offriront trois modes d'identification des ressources : une navigation par collection, une interface simple de recherche dans les métadonnées et une interface complexe de recherche à facette.