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

Un équipement d'excellence de mutualisation de ressources et d'outils sur la langue et son traitement informatique

Jean-Marie Pierrel

ORTOLANG bénéficie d'une aide de l'Etat au titre du programme « Investissements d'avenir » (ANR-11-EQPX-0032)

Pourquoi un tel équipement ?

- Pour servir de **support aux travaux de recherche** :
  - La notion de corpus est aujourd'hui incontournable spécifiquement en linguistique et en TAL ;
  - sans une véritable mutualisation chaque équipe de recherche se verrait dans l'obligation de tout réinventer, alors même que nul ne peut être spécialiste de chacun sous-domaine.
- Pour la **valorisation des résultats de recherche (corpus, lexique, dictionnaires et outils de traitement)**
  - Un des aspects essentiels aujourd'hui est leur informatisation et leur disponibilité sur la toile sous une forme facilement accessible et exploitable par l'ensemble de la communauté scientifique et industrielle.

Une nécessité : assurer le partage et la pérennisation de ressources

- Un rôle central dans de nombreux domaines de recherche en linguistique et en TAL
- Nécessaires pour l'émergence et l'apprentissage de modèles
  - pour les approches stochastiques
  - pour les approches symboliques
- Indispensables pour valider et évaluer théories et outils
- Nécessité de corpus de référence ou « étalons » pour évaluer nos productions de recherche

ORTOLANG un équipement en réseau réunissant des compétences complémentaires

Une réseau autour de trois pôles géographiques

- **Nancy**
  - **ATILF** (Analyse et Traitement Informatique de la Langue française)
    - **CNRTL** (centre national de ressources sur l'écrit)
  - **LORIA** (Pôle TALC)
  - **INIST** (soutien centre Clarin et métadonnées ; hébergeur de l'infrastructure informatique)
- **AIX en Provence**
  - **LPL** (Laboratoire Parole et Langage)
    - **SLDR** (Speech & Language Data Repository)
- **Paris Nanterre et Orléans**
  - **MODYCO** (Modèles, Dynamiques, Corpus)
  - **LLL** (Laboratoire Ligérien de Linguistique**)**

Un centre de diffusion

- **Dépôt fiable de données et entrepôt OAI-PMH permettant :**
  - Une identification pérenne des ressources (Handle).
  - Une preuve d'intégrité de la donnée associée à un identifiant pérenne
  - La gestion de versions.
  - L'authentification des utilisateurs lors de l'accès à des données à accès restreint.
  - La mise à disposition d'outils de traitement de données sous forme de Web Services.

Trois grandes « thématiques »

- **Ecrit** autour du CNRTL (Nancy)
- **Oral** autour du SLDR (Aix)
- **Patrimonialisation** autour de Paris-Orléans
- Centres thématiques : parties émergées de l'équipement
  - s'appuyant sur le centre de diffusion commun
  - directement visibles pour les utilisateurs
  - fortement connectés,
  - présentant des centres d'intérêt différents
    - méthodes de navigation et de recherches différentes,
    - outils différents

Avec quels objectifs (1) ?

- **Identification/préparation des données :**
  - **catalogage** des ressources et outils existants à travers un ensemble de métadonnées normalisées ;
  - **contrôle et validation** des ressources et des outils : accompagnement des auteurs sur les standards, les normes et les recommandations internationales actuelles : XML, TEI, LMF, MAF et SYNAF ;
  - **enrichissement** de ressources et des outils.

Avec quelles objectifs (2) ?

- **Archivage :**
  - **stockage, maintenance et curation** des ressources et des outils ;
  - **archivage** pérenne, à travers la solution mise en place par la TGIR Huma-Num (fusion d'ADONIS et de CORPUS) en lien avec le CINES.
- **Diffusion :**
  - **aide et accompagnement des utilisateurs** et mise en place des procédures permettant à des utilisateurs de la plateforme d'exploiter les ressources et outils mutualisés sans avoir à se soucier de leur localisation et implantation géographiques.

Objectifs de la phase de mise en place (2013-2016)

- Mise en place de l'architecture matérielle et logicielle (cf. ci-après)
- Développement de ressources et d'outils de base
  - Viewer de corpus et d'annotations
  - Annotation morphosyntaxique pour l'écrit
  - Détection d'entités nommées
  - Etc..
- Enrichissement de Corpus
  - Standardisation et mise à disposition de nouvelles ressources

Architecture logicielle

- Un serveur de diffusion
  - Stocke toutes les données et les métadonnées
  - Moissonnable en OAI-PMH par tout le monde
  - Accès direct aux documents par tout le monde via un identifiant pérenne
  - Organise l'archivage pérenne (vers le CINES)
  - Web Services de recherche dans les métadonnées et les données textuelles
- Les serveurs thématiques
  - Offrent des interfaces de recherche et de visualisation qui leur sont spécifiques (Ecrit, Oral, patrimonialisation)
  - Proposent des ressources téléchargeables et outils pour exploiter ces ressources
  - Propose un espace de travail temporaire pour enrichir les ressources avant diffusion

Infrastructure

- L'infrastructure
  - Hébergement du matériel à l'INIST : on bénéficie des moyens et de l'expérience de l'INIST
    - Réseau, Sécurité, Salles serveurs, Exploitation
  - Architecture virtualisée (une dizaine de machines virtuelles) sur un cluster de 3 serveurs DELL R620 Bi-processeurs / 8 cœurs, 128 Go RAM par serveur
  - Continuité de service et Haute disponibilité
  - Jouvence des équipements prévue en 2016 et 2019
- Stockage
  - Une surface disque dédiée de 40 To « utiles » en RAID 6 (= 18 x 3 To brut)
- Sauvegarde
  - Utilisation de l'infrastructure de sauvegarde de l'INIST, avec pool de supports dédié

Planning

- Lancement des procédures d'acquisitions (02-03/2013)
- Mise en place technique des machines, des systèmes d'Exploitation et de Virtualisation, couplage avec l'infrastructure de stockage, livraisons aux équipes en charge de la mise au point de la plateforme (03-05/2013)
- Livraison des environnements (06/2013)
- Premières ressources diffusées (3ème semestre 2013)

Quelques caractéristiques d'ORTOLANG

- Une ouverture **pluridisciplinaire**
  - SHS & Informatique
- **S'appuyant sur l'existant** des centres nationaux de ressources
  - CNRTL et SLDR
- **Gérant des ressources pour l'ensemble de la communauté scientifique**
  - ORTOLANG n'est qu'une infrastructure de mutualisation : les Corpus et les outils restent propriété des laboratoires
  - Nous avons prévu des moyens pour aider des laboratoires à finaliser et normaliser leurs ressources
- **Droits d'accès définis par les propriétaires des corpus mais recommandations d'ORTOLANG :**
  - Liberté d'usage pour la recherche et tant qu'il n'y a pas de *business*
  - Moyennant royalties dès qu'il y a du *business*

Insertion dans le dispositif national (1)

- Liaison avec la **TGIR Huma-Num**
  fusion d'ADONIS et de CORPUS
  - JM Pierrel (ATILF) membre du comité de pilotage du Consortium Corpus ECRITS
  - G. Bergounioux (LLL), Christophe Parisse (Modyco) membres du comité de pilotage du Consortium IRCOM
  - Historiquement le CNRTL et le SLDR, opérateurs d'ADONIS dans leur domaine => ORTOLANG opérateur d'Huma-Num pour le domaine linguistique
    - Ortolang : service spécialisé complémentaire de la grille de calcul d'Huma-Num
    - Archivage pérenne assuré par Huma-Num
  - Appels communs à projets envisagés pour la standardisation de Corpus

Insertion dans le dispositif national (2)

- Liaison avec la **fédération ILF** (Institut de linguistique Française)
  - ORTOLANG focalisé sur la Langue Française

    => complémentaire de la fédération ILF
  - Plateforme d'accueil pour le projet « corpus de référence du français »
- Liaison avec la **fédération TUL** (Typologie et Universaux linguistiques)
  - Ouvert vers l'accueil et la gestion des corpus des chercheurs français en linguistique

Insertion dans le dispositif international

- **DARIAH :** Digital Research Infrastructure for the Arts anh Humanities
  - Participation active de ses laboratoires supports
- **CLARIN :** Common Language Resources and Technology Infrastructure
  - Objectif : Stabiliser un réseau français compatible avec l'infrastructure CLARIN
  - Rappel : l'ATILF partenaire du projet européen de définition de CLARIN

Relations avec les partenaires extérieurs

- Divers **contacts avec des partenaires externes** ayant déposé ou souhaitant **déposer leurs ressources sur ORTOLANG**
- Vers des appels commun entre ORTOLANG et les consortiums d'Huma-Num en linguistique
  - corpus sur les **nouvelles-formes de communication** en français dans le cadre du GT7 du consortium Ecrit de **CORPUS-IR**.
- Support pour différents nouveaux projets :
  - Mise en place de l'initiative "Corpus de référence du français »
  - Projet d'annotation d'entités nommées (avec le LI de Tours)
  - **Projets ANR** (ORFEO, OTIM, etc.)
- **Labex EFL (Paris), BLRI (Aix-Marseille)**

Quelle structure ?

- **Directeur** : Responsable scientifique et technique
  - J.M. Pierrel
- Un **comité technique opérationnel**
  - composé d'un représentant de chaque partenaire
- Un **comité scientifique**
  - largement ouvert à l'international
- Un **comité d'orientation**
  - regroupant les représentants des tutelles des partenaires

Quel moyens ?

- **Investissement** : 2 200 K€ (2013-2016)
  - Dont 1 530 K€ de frais de personnel pour l'élaboration de
    - l'infrastructure logicielle,
    - des ressources
    - des outils
- **Fonctionnement** : 400 K€ (2016-2019
- **TOTAL : 2 600 K€ HT**
- **Signature de la convention : 15 janvier 2013**

Pour en savoir plus et suivre l'avancement du projet

www.ortolang.fr