

Présentation de la plateforme DeepCorp adaptée à ESLO



Laboratoire
Ligérien de
Linguistique

BRAS Enzo - année 2023
ASTN/LLL/CNRS



Sommaire

- Les besoins
- Analyse de l'existant
- Nouveau site
 - ◆ CouchBase
 - ◆ Spring Boot
 - ◆ Wordpress
- Amélioration / Projection
- Bilan
 - ◆ Problèmes rencontrés
 - ◆ Ressenti



Les besoins



Le services ASTN a entrepris le développement de plusieurs plateformes d'outillage. Une des premières plateformes développées est AlimCorp. Alimcorp couvre les phases de création, de collecte et d'enrichissement de Corpus.

- La solution actuelle web du site d'ESLO qui permet aux chercheurs de naviguer dans le Corpus est devenue obsolète et ne permet pas d'évolution majeure.
- La solution retenue par le service ASTN est donc le développement d'une plateforme dédiée à l'extraction et aux traitements sur les données de ces corpus, plateforme nommée DeepCorp.
- Nouvelle plateforme qui doit permettre d'offrir des fonctionnalités de traitements et d'analyses de haut niveau aux chercheurs à partir des données collectées par AlimCorp.

Site associé au Corpus Eslo

perte de la maitrise

joomla



The screenshot shows a Joomla-based website for ESLO (Enquêtes Sociolinguistiques à Orléans). The browser address bar displays "eslo.huma-num.fr/index.php". The website features a blue header with the ESLO logo and navigation menu items: Accueil, Présentation, Le corpus, Méthodologie, and La recherche. The main content area includes a welcome message, a description of the linguistic corpus, and contact information. The footer contains logos for CIRS, Centre-Voi de Loire, NR, and Administration with a Connexion button.

eslo.huma-num.fr/index.php

eslo
Enquêtes Sociolinguistiques à Orléans

Accueil Présentation Le corpus Méthodologie La recherche

Bienvenue sur le site du portrait sonore d'Orléans (ESLO)

« Accueil
Présentation
« Eslo en quelques mots
« Le projet scientifique
Le corpus
« Présentation du corpus
« Accéder au corpus
Méthodologie
La recherche
« L'équipe
« Projets de l'équipe et sous corpus
« Projets Partenaires
« Evénements

« Ce site met à votre disposition un corpus linguistique composé d'enregistrements sonores et de leurs transcriptions réalisés à Orléans entre 1968 et 1974 (ESLO1) et à partir de 2008 (ESLO2).
[En savoir plus...](#)

« Accéder à l'ensemble des enregistrements et transcriptions : [consulter et télécharger les documents](#) (la base de données est en cours d'alimentation).

« Tous les Orléanais ont la parole : Venez [participer](#) au portrait sonore de la ville d'Orléans par ses habitants !

Courriel : eslo.lsh@univ-orleans.fr
Téléphone : 02 38 49 40 10

CIRS Centre-Voi de Loire NR Administration Connexion

Difficulté à faire évoluer

Page formulaire de saisie de critère de recherche

Accès restreint La majorité des documents sont en libre accès, le reste du corpus peut-être consulté par des chercheurs après signature d'une convention spécifique. La demande de convention doit être faite par mail à eslo.lsh@univ-orleans.fr

Corpus ESLO > Accéder au corpus > Recherches dans les catalogues

Recherches dans les catalogues

Si aucun corpus n'est sélectionné, tous les corpus seront interrogés.

Sélection du corpus

ESLO1 ESLO2

Sélection catégorie(s)

Entretien	Entretien
Contact	Repas
Ouverture de l'entretien	Interview de personnalités
Cloture de l'entretien	Conférences
Repas	Itinéraire
Magasin	24heures

▼ ▲ ▼ ▲

▼ ▲ ▼ ▲

Veillez sélectionner le catalogue de recherche :

Enregistrement

Pour prendre en compte les critères de recherche, vous devez cocher la ligne concernée.

Enregistrement	Titre	Date	Spatial	Acoustique
<input type="checkbox"/>	<input type="text"/>	du <input type="text"/> au <input type="text"/>	<input type="text"/>	<input type="text"/>
<input type="checkbox"/>				
<input type="checkbox"/>				
<input type="checkbox"/>				Mauvaise

Rechercher Réinitialiser

Affichage résultat d'une recherche

Accès restreint La majorité des documents sont en libre accès, le reste du corpus peut-être consulté par des chercheurs après signature d'une convention spécifique. La demande de convention doit être faite par mail à eslo.lib@univ-orleans.fr

Corpus ESLO > Accéder au corpus > Recherches dans les catalogues

Résultat(s)

Descendre jusqu'au formulaire Réinitialiser la recherche

Enregistrement(s)
ESLO1_ENT_001
ESLO1_ENT_002
ESLO1_ENT_003
ESLO1_ENT_004
ESLO1_ENT_005
ESLO1_ENT_006
ESLO1_ENT_007
ESLO1_ENT_008
ESLO1_ENT_009
ESLO1_ENT_010
ESLO1_ENT_011
ESLO1_ENT_012
ESLO1_ENT_013
ESLO1_ENT_014
ESLO1_ENT_015

809 enregistrements trouvés sur un total de 809 (tous les corpus) << Page 1 sur 54 >>

Remonter au début des résultats

Recherches dans les catalogues

Si aucun corpus n'est sélectionné, tous les corpus seront interrogés.

Sélection du corpus

ESLO1 ESLO2

Sélection catégorie(s)

Entretien Entretien

Accès aux détails d'un enregistrement

eslo-huma-num.fr/CorpusEslo/hml/fiche/enregistrement?id=327

Accès restreint

La majorité des documents sont en libre accès, le reste des corpus peut être consulté par des chercheurs après signature d'une convention spécifique. La demande de convention doit être faite par mail à eslo.huma@univ-orleans.fr

Accéder au corpus » Recherches » Consultation des fiches de catalogue » Fiche enregistrement

Fiche enregistrement

Référence enregistrement: ESLO1_CONF_503

Fichier son: ESLO1_CONF_503.wav

Corpus: ESLO1

Catégorie: Conférences

Précisions sur la catégorie: Enregistrements de conférences suivies éventuellement de discussions

Sujet: (text_and_corpus_linguistics) Français (Ethnologue: fra)
Enseignement ; conférences/conférence-discussion ; l'enseignement des langues modernes/étranger ; historique ; école de la méthode directe ; nouvelle orientation psychopédagogique ; site FRESS & PIAGET ; Traité de psychologie expérimentale ; SINCLAIR DE ZWAART ; l'Acquisition du langage chez l'enfant ; MALAURET ; Psychopédagogie des méthodes audiolinguales. Compétence de l'auditeur/élève ; perception globale, sélective ; IR3 sémantique ;

Editeurs: LLL Université d'Orléans

Créateurs: LLL Université d'Orléans - ESLOs

Chercheurs:

- Blanc, Michel
- Biaggi, Patricia

Chercheurs locuteurs:

Participants:

Description des participants:

Descriptions annexes:

Remarques: Chercheur ; non-enseignant

Fiche modifiée par: oboisde

Date d'enregistrement: 26/02/1970

Droits: Copyright (c) 2012 Université d'Orléans/LLLFreely available for non-commercial use. This file is licensed under a Creative Commons License.

Format: (IANA MIME Media Type: audio/wav)

Durée: 01:35:00

Accoustique: Excellente

Précisions acoustiques: conférence très claire ; discussion médiane (enchevêtrement de voix)

Lieu spatial: Orléans

Lieu TGN: 7008337

Lieu Point: east=1.96, north=47.90

Locuteurs:

- 503L.OC1
- 503L.CONF
- 503L.OC2
- 503L.OC3
- 503L.OC4
- 503L.OC5
- 503L.OC6
- 503L.OC9
- 503L.OC10
- 503L.OC11
- 503L.OC12

Transcriptions:

- ESLO1_CONF_503_A
- ESLO1_CONF_503_B
- ESLO1_CONF_503_C

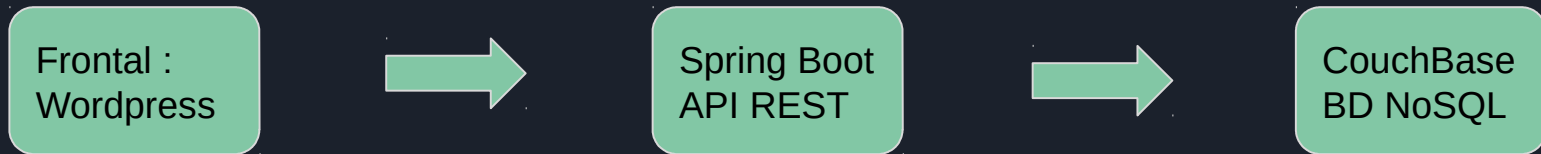
- Télécharger les métadonnées au format XML.
- Télécharger le fichier sonore
- Jouer l'enregistrement avec la dernière transcription
- Revenir au formulaire d'accès au catalogue
- Revenir au formulaire de requête dans le corpus
- Revenir sur la page d'index des accès aux corpus

Nouveau site DeepCorp

Projet : prototyper une solution d'extraction et d'analyse à partir du modèle Eslo.

La solution s'appuie sur une architecture multi-tiers constituée :

- Frontal sous wordpress
- Web service (spring boot)
- Backend : couchBase et Kubernetes



Choix du CMS était imposé par le service ASTN.

Backend / Spring Boot

- Ce web service a pour objectif de desservir et d'alimenter le frontal en effectuant la demande d'exécution des requêtes au cluster CouchBase. Il a pour objectif de mutualiser les demandes car ASTN a plusieurs projets qui ont besoin de ce service. De plus, le fait de séparer la logique de stockage du frontal permet le cas échéant de changer de solution si besoin.

```
1 package fr.ill.refcorp;
2
3 import java.util.ArrayList;
4
5 @RestController
6 public class DataControllerEnregistrements {
7
8     @Autowired
9     private Cluster CouchbaseServerPersistentProvider;
10
11     @Autowired
12     private Environment env;
13
14     /**
15      * Retourne tous les enregistrements et leur contenu
16      * @return : le contenu de tous les enregistrements
17      */
18     @GetMapping("/get_enregistrements")
19     public List<Map<String, Object>> get_enregistrements() {
20         QueryResult result = CouchbaseServerPersistentProvider.query("SELECT * FROM enregistrements");
21         List<Map<String, Object>> resultItems = new ArrayList<>();
22         for (JsonObject row : result.rowsAsObject()) {
23             System.out.println("row: " + row);
24             resultItems.add(row.toMap());
25         }
26         return resultItems;
27     }
28
29     /**
30      * Retourne toutes les enregistrements (le contenu)
31      * @param body
32      * @return
33      */
34     @GetMapping("/get_enregistrements_contenu")
35     public Map<String, Map<String, Object>> get_enregistrements_contenu() {
36         QueryResult result = CouchbaseServerPersistentProvider.query("SELECT META(id) * FROM enregistrements t1");
37         Map<String, Map<String, Object>> resultItems = new TreeMap<>();
38         for (JsonObject row : result.rowsAsObject()) {
39             String key = (String) row.get("id");
40         }
41     }
42 }
```



Travail à effectuer

- Mon objectif est de développer principalement la partie Extraction, sachant que les autres pages ne sont que informations statiques que l'on complétera par la suite.
- Pour la partie extraction, je vais reproduire l'existant et eprmettre des extractions sur les types de données enregistrements, transcriptions et locuteurs.
- Je vais ensuite rajouter les possibilités de traitements aux autres types de données de type segments et d'offrir la possibilité d'explorer également les descriptions plus précises des segments.

Frontal / page formulaire

formulaire recherche

Sélection du corpus

ESLO1 ESLO2

Sélection catégorie(s)

Ouverture de l'entretien Cloture de l'entretien Repas Magasin Divers	Entretien Repas Interview de personnalités Conférences
--	---

▼ ▲ ▼ ▲

▼ ▲ ▼ ▲

Vous devez sélectionner le catalogue de recherche :

Locuteur ▼

Pour prendre en compte les critères de recherche, vous devez cocher la ligne concernée.

Locuteur

<input type="checkbox"/> Référence	<input type="text"/>
<input type="checkbox"/> Sexe	Masculin <input type="radio"/> Féminin <input type="radio"/>
<input type="checkbox"/> Tranche d'Age	- de 5 ans ▼
<input type="checkbox"/> Catégorie Professionnelle (INSEE)	Ouvriers ▼
<input type="checkbox"/> Niveau d'études	À définir ▼

Rechercher Réinitialiser

formulaire recherche

Sélection du corpus

ESLO1 ESLO2

Sélection catégorie(s)

Contact Divers Interview de personnalités Conférences	Entretien Repas Interview de personnalités Conférences
--	---

▼ ▲ ▼ ▲

▼ ▲ ▼ ▲

Vous devez sélectionner le catalogue de recherche :

Segment ▼

Pour prendre en compte les critères de recherche, vous devez cocher la ligne concernée.

Segment

<input type="checkbox"/> Enregistrement	<input type="text"/>
<input type="checkbox"/> Debut / Fin	entre <input type="text"/> et <input type="text"/>
<input type="checkbox"/> Texte	<input type="text"/>
<input type="checkbox"/> Locuteur	<input type="text"/>
<input type="checkbox"/> Format	Ex: PRON VERB <input type="text"/>

Rechercher Réinitialiser

Nouveau site / page formulaire

Locuteur

Référence

Sexe Masculin Feminin

Tranche d'Age -de 5 ans

Catégorie Professionnelle (INSEE) Ouvriers

Niveau d'études À définir

RM136

HV440

TK473

HZ912

HS609

VP283

LF422

TS894

ZZ876

DT708

FH253

FT070

KF467

CW903

GM125

GX031

JZ001

ZZ876

CATÉGORIE SOCIAL : Autres personnes sans activités prof.

REMARQUE :

[détail](#)

Enregistrement

Titre

Date du / / au / /

Spatial

Acoustique Bonne

ESLO1_CONF_501

ESLO1_CONF_502

ESLO1_CONF_503

ESLO1_CONF_504

ESLO1_CONF_505

ESLO1_CONSCMPP_701

ESLO1_CONSCMPP_702

ESLO1_CONSCMPP_703

ESLO1_CONSCMPP_704

ESLO1_CONSCMPP_705

ESLO1_CONSCMPP_707

ESLO1_CONSCMPP_708

ESLO1_CONSCMPP_709

ESLO1_CONSCMPP_710

ESLO1_CONSCMPP_711

ESLO1_CONSCMPP_712

ESLO1_CONF_504

SOMMAIRE : Enseignement: conférencesconférence: la mathématique moderne et les problèmes d#évaluation.résumé-bilan général de la pédagogie en mathématiques; la docimologie; importance des barèmes; compte-rendu d#expérience; une copie soumise à 30 professeurs pour vérifier écarts d#évaluation avec / sans barème; faible rendement d#étudiants en mathématiques; besoin croissant de scientifiques; nécessité de rénover programmes de mathématiques.

DURÉE : 97

[détail](#)

ESLO1_CONSCMPP_705

SOMMAIRE : enregistrement : présentation du problème

DURÉE : 18

[détail](#)

Nouveau site / page détail



La fiche enregistrement

Statut	
Description_enregistrement	
Droits	Copyright (c) 2012 Université d#Orléans/LLL
Langue	
Locuteurs	504CONF 504LOC1 504LOC2 504LOC3 504LOC4 504LOC5 504LOC6 504LOC11 504LOC15
Precisioncategorie	Enregistrements de conférences suivies éventuellement de discussions
Acoustique	2
Precision_acoustique	claire mais un peu sourde
Participants	
Created	1970-03-05
Categorie	Conférences
File_enreg	0

Nouveau site / page comparaison



Comparaison

Enregistrement ▼ ESLO1_CONF_504

Statut		
Description_enregistrement		
Droits	Copyright (c) 2012 Université d#Orléans/LLL	Copyright (c) 2012 Université d#Orléans/LLL
Langue		
Locuteurs		504CONF 504LOC1 504LOC2 504LOC3 504LOC4 504LOC5 504LOC6 504LOC11 504LOC15
Precisioncategorie	Enregistrements de conférences suivies éventuellement de discussions	Enregistrements de conférences suivies éventuellement de discussions
Acoustique	5	2
Precision_acoustique	m2DIOCRE	claire mais un peu sourde
Participants		
Created	1970-01-31	1970-03-05
Categorie	Conférences	Conférences
File_enreg	0	0
Format	0	0
Chercheurs	47 Blanc Michel 51 Biggs Patricia	47 Blanc Michel 51 Biggs Patricia

Suivi et tuning / logs

```
mysql> select * from wp_logs_exec;
+-----+-----+-----+
| user_ID | catalogue | requete |
+-----+-----+-----+
|         |           |         |
+-----+-----+-----+
| 1 | enregistrement | ( Modele = 'Eslo1' ) AND ( Enregistrement.Categorie = 'Entretien' OR Enregistrement.Categorie = 'Magasin' ) AND Enregistrement.Acoustique = 4 |
|         |           |         |
| 1 | locuteur | ( enregistrements.Modele = 'Eslo1' OR enregistrements.Modele = 'Eslo2' ) AND locuteur.Sexe = 1 AND locuteur.Tranche like '40/45' AND locuteur.categ_socio like 'Ouvriers' |
|         |           |         |
| 1 | transcription | ( enregistrements.Modele = 'Eslo1' OR enregistrements.Modele = 'Eslo2' ) |
|         |           |         |
| 1 | segment | 0=0 |
|         |           |         |
| 1 | enregistrement | ( Modele = 'Eslo1' ) AND ( Enregistrement.Categorie = 'Reunion' OR Enregistrement.Categorie = 'Repas' OR Enregistrement.Categorie = 'Appel téléphonique' ) AND Enregistrem |
|         |           |         |
| 1 | enregistrement | ( Modele = 'Eslo1' OR Modele = 'Eslo2' ) AND Enregistrement.Identifiant like '%ESL01_CONF_505%' |
|         |           |         |
+-----+-----+-----+
6 rows in set (0.00 sec)
```

Afin d'offrir un service de qualité, un système de mise en place de logs permet de notifier les requêtes effectuées et leur temps d'exécution.

Ce relevé de métriques va nous permettre de suivre la capacité du backend à répondre aux différentes sollicitations.

Amélioration / Projection

- Ergonomie
- Gestion des erreurs
- Ajout de graphique dans la page comparaison
- Faire un tri dans les éléments affichés dans la page détail



Bilan / Problème rencontré

- développement de l'API REST
- compréhension de WordPress



Bilan / Ressenti

- développement de nouvelles compétences
- des choses à améliorer



Merci pour votre écoute

WAV

MP3

DeepCorp

ESLO

IA

API (next)

Kubernetes

20%

25%

35%

20%

$$r_{j,m} = \frac{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{im} - \bar{x}_m)}{s_j \times s_m}$$