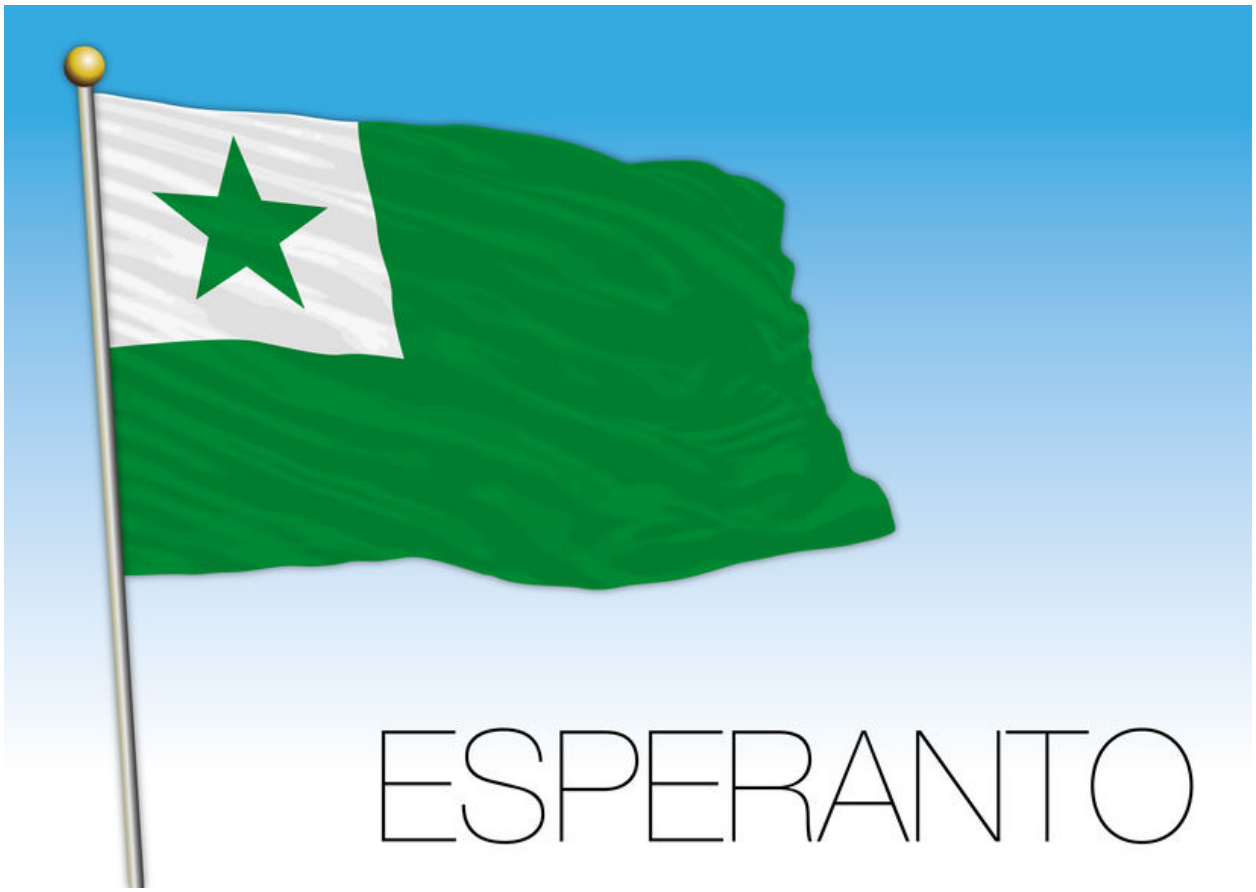
Cahier des charges

Construction d'une base de données d'oeuvres culturelles en espéranto

Novembre à Juin 2021

Alexandre DURY 11919722



Contexte du projet

L'**espéranto** est une langue internationale parlée à travers plus de **120 pays** dans le monde, par quelque **deux millions** de personnes. Les recherches effectuées sur le sujet démontrent d'ailleurs que de plus en plus de personnes le parlent, notamment grâce aux cours proposés sur internet.

Comme cette langue n'est rattachée à aucun État, l'une de ses critiques porte sur sa **limitation au niveau culturel**. Pourtant, de nombreuses œuvres (romans, musique, films et vidéos, magazines, etc.) sont disponibles en espéranto, qui peuvent être soit des traductions d'œuvres existantes (la eta princo pour le petit prince, 700 limerikoj, etc...) soit des oeuvres originales de littérature espérantophone (Mr. Tot aĉetas mil okulojn). À ce jour, il n'existe pas de liste la plus exhaustive possible rassemblant ces œuvres espérantistes. C'est pour cette raison que nous souhaitons concevoir une base de données regroupant un grand nombre d'œuvres dans cette langue.

Pour vous donner un aperçu, voici à quoi ressemble l'espéranto :

Native Esperanto speaker | Stela speaking the Esperanto langu…

Dans le cadre de ce projet, Monsieur Fabien Duchateau, maître de conférences au LIRIS (Laboratoire d'Informatique en Image et Systèmes d'information) et membres de l'équipe « Bases de données », propose le sujet de projet POM (Projet d'Orientation en Master) intitulé « Construction d'une base de données d'oeuvres culturelles en espéranto ».

Objectif

L'objectif général de ce projet POM sera de regrouper des œuvres en espéranto au même endroit, et idéalement de les enrichir par des liens vers des concepts connexes. Plusieurs sous-objectifs sont identifiables.

I. Modélisation de la base de données

Ce projet a pour objectif le développement d'une base de données recensant un maximum d'œuvres en espéranto (livres, poèmes, musique, films et vidéos, presse, radio, sites web, etc.). La première étape consiste donc à modéliser les besoins. Ensuite, les œuvres seront récupérées principalement sur plusieurs sources de données, en particulier des sites web (BNF, UEA katalogo, ANL, Open Library) afin de collecter un maximum d'œuvres ainsi que d'informations les concernant.

II. Intégration et enrichissement

Les multiples œuvres seront collectées en utilisant les API des différentes sources de données ou en "scrapant" le contenu des documents HTML. Comme une œuvre peut être présente sur plusieurs sources, il est nécessaire d'implémenter un algorithme d'intégration pour détecter si une œuvre existe déjà dans la base et si ses informations sont complètes. Pour enrichir les données intégrées, les œuvres et auteur(e)s seront reliées à d'autres bases de connaissances (VIAF, BNF, **World cat**, DBpedia, ...) . Ce processus appelé "entity linking" permet par exemple d'avoir davantage d'informations sur l'auteur ou de découvrir les versions d'une œuvre en langue originale.

III. Présentation des oeuvres

Un site web sera développé pour visualiser les œuvres espérantistes intégrés à la base de données, avec une fonctionnalité permettant la recherche des œuvres culturelles. Si le temps le permet, une demande d'insertion de nouvelles œuvres, ou encore une demande de modification liée à une œuvre présente sur le site, pourra être développée.

Contraintes

- L'application doit récupérer des données existantes via une API ou autre, à partir de sources externes: UEA katalog et BNF. Le processus doit pouvoir être relancé afin de collecter d'éventuelles nouvelles œuvres.
- Lors de la saisie d'une nouvelle œuvre, Il faut s' assurer au préalable qu'elle n'existe pas déjà dans la BD.
- Prise en compte du multilinguisme.
- Site Web en PHP/HTML/CSS.
- Utilisation du SGBD MySQL.

Travail à réaliser

Afin d'implémenter les besoins du chapitre précédent, nous allons respecter la démarche suivante.

• Modélisation de la base de données

Dans un premier temps nous modélisons la base de données à l'aide d'outils (Mocodo, Looping) pour partir d'une structure correcte et ainsi insérer toutes les données nécessaires, associées à une œuvre. On pourra s'inspirer des ontologies existantes dans le domaine bibliographique qui décrivent déjà des œuvres (BIBFRAME, LD4L, RDA).

• Implémentation d'algorithmes d'intégration

L'algorithme d'intégration peut se faire de deux manières, à la volée(lorsqu'on intègre une nouvelle source) ou après avoir récolté toutes les données (déduplication). Dans un premier, temps nous utiliserons le système d'implémentation à la volée puis, au fur et à mesure nous exploiterons la méthode de factorisation, afin de réduire au maximum les duplicatas s'ils existent.

• Implémentation d'un algorithme pour la détection de liens vers d'autres référentiels

Certaines œuvres manqueront d'informations comme par exemple un résumé, un prix, … Ces informations manquantes pourront être ajoutées grâce à la détection de liens vers d'autres ressources (VIAF, BNF, **World cat**, DBpedia, ...). L'algorithme cherchera les entités en combinant plusieurs de ses caractéristiques (e.g., titre + année).

• **Développement d'un site web pour visualiser les données.**

Tout d'abord nous développerons un site qui affichera les données regroupées dans la base créée, puis si le temps le permet, nous ajouterons un maximum de fonctionnalités liés au projet, comme par exemple la demande d'ajout d'une œuvre non présentée sur le site, mais aussi une demande de modification si des informations affichées semblent incorrectes. La possibilité de connexion avec 2 types d'utilisateurs, les utilisateurs lambdas qui eux pourront faire les demandes citées ci-dessus, et un administrateur qui pourra traiter ces demandes, c'est-à-dire qu'il aura plus de droit sur le site, grâce à des fonctionnalités qui lui sont propres.

Voici un **tableau de répartition du travail** à effectuer :

Mi-Novembre	Modélisation du MLR de la base de données
Fin-Novembre	Ajout/correction du MLR
Début-Décembre	Vérifier/ compléter du script de création
Mi-Décembre	Compréhension de l'API de la source utilisé
Fin-Décembre	Test API avec récupération de données basiques
Janvier	Script de récupération automatique de données
Février	Détection de liens vers d'autres référentiels
Mars	Développement du site pour visualiser/rechercher les oeuvres
Avril	Ajout de fonctionnalité sur le site
Mai-Juin	Finalisation des fonctionnalités et rapport