Projet d'Orientation en Master

Contexte

Langue espéranto :

- Créée par le Docteur Zamenhof en 1887
- 2 à 7 millions de locuteurs
- Oeuvres nationales traduites et oeuvres originales en espéranto
- Limitation culturelle

Oeuvres

- Nombreuses oeuvres espérantistes, mais pas de liste exhaustive
- Difficulté pour trouver une oeuvre

⇒ Construction d'une base de données d'oeuvres culturelles en lien avec l'espéranto

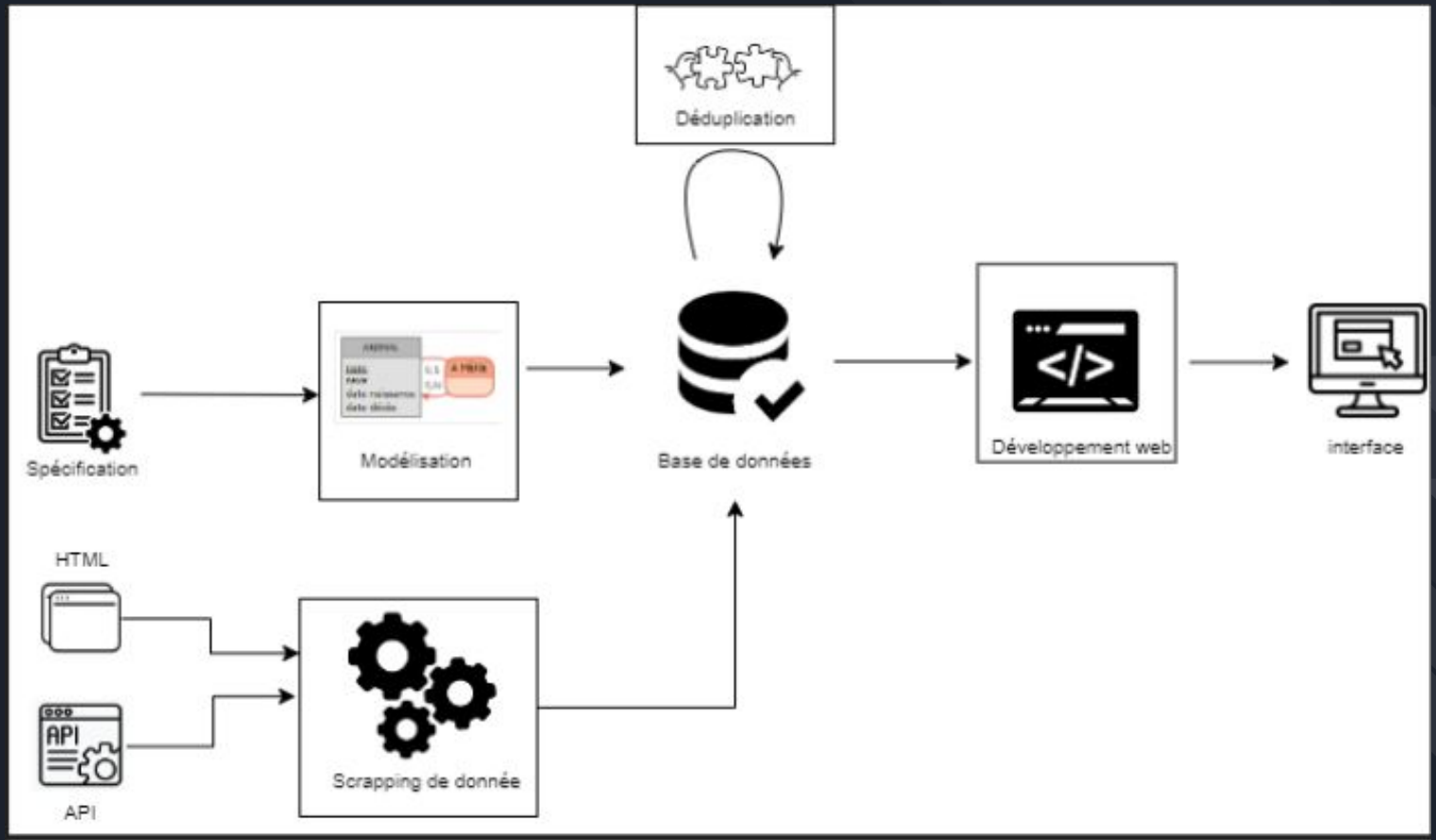
Problématiques

- Comment assembler des informations issues de différentes sources de données ?

- Comment limiter les redondances dans ces données, en particulier pour les œuvres et agents ?

- Comment présenter ces informations ?

Proposition

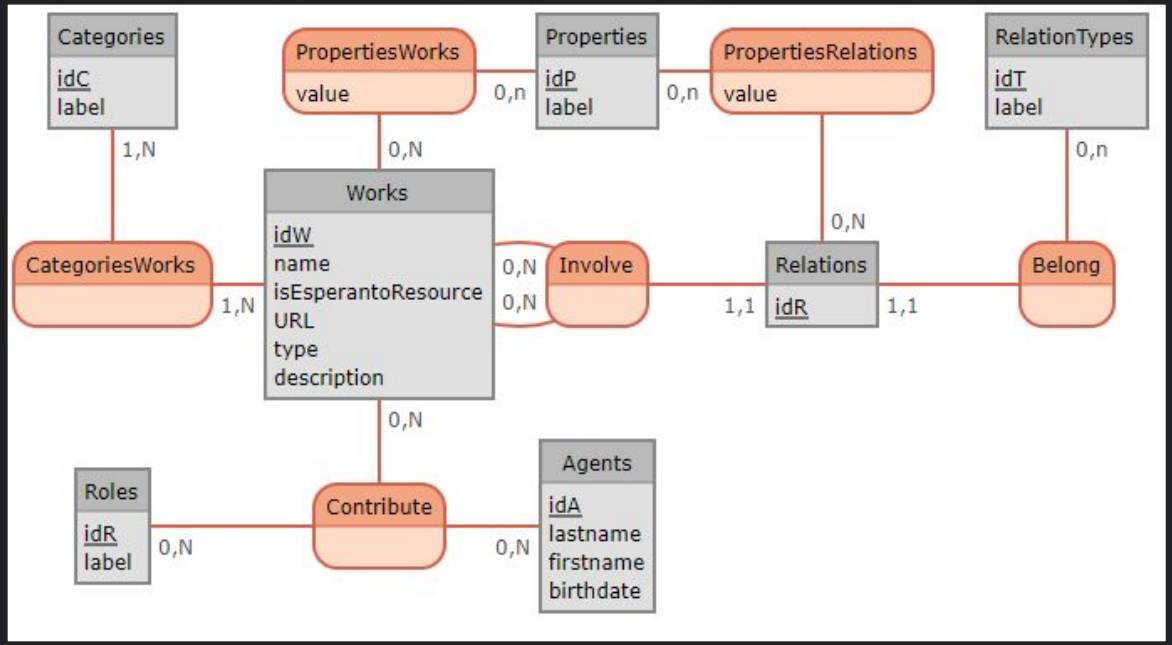


Sommaire

A. Modélisation de la base de donnée

MCD online

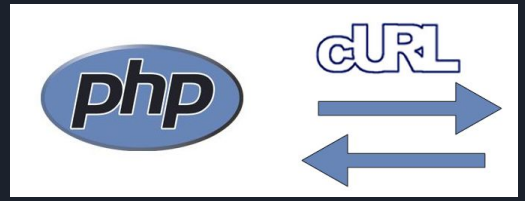
```
Categories: idC[int(10)], label[varchar(100)]
PropertiesWorks, 0n Properties, 0N Works: value[varchar(100)]
Properties: idP[int(10)], label[varchar(200)]
PropertiesRelations, 0n Properties, 0N Relations: value[varchar(100)]
RelationTypes: idT[int(10)], label[varchar(200)]
```



B. Scraping de données

OpenLibrary

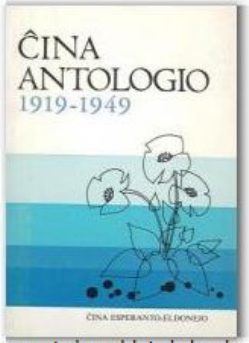

```
"description": {
    "type": "/type/text",
    "value": "Klasika, facile legebla prezento de la trajtoj de
        Esperanto el moderna scienca vidpunkto. La fonetiko,
        morfologio, sintakso, leksiko kaj semantiko estas klarigitaj
        per multaj ekzemploj el aliaj lingvoj."
},
"covers": [
    7105333
],
"last_modified": {
    "type": "/type/datetime",
    "value": "2012-04-28T12:22:02.605320"
},
"latest_revision": 5,
"key": "/works/OL4329057W",
```

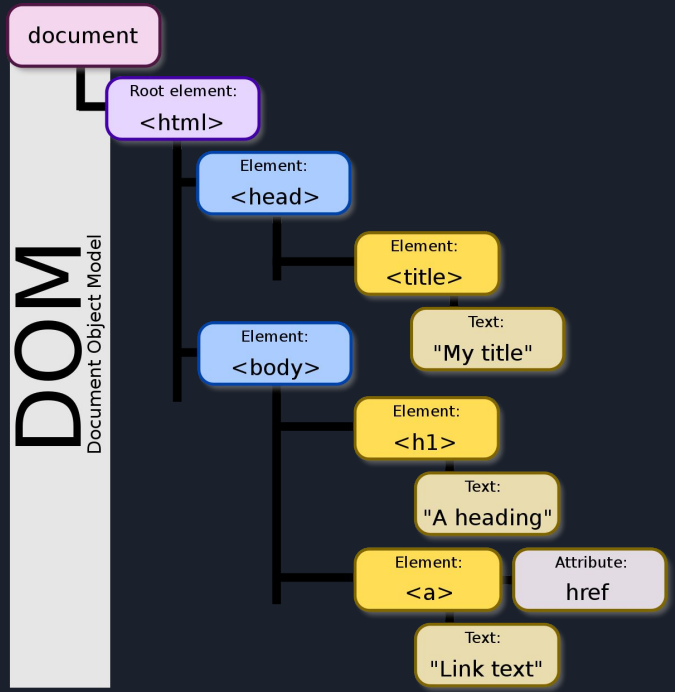
B. Scraping de données

Katalogo

Titolo	Ĉina antologio (1919-1949)
Kategorio	Antologioj
Prezo	Nehavebla
Eldonloko, jaro	Pekino, 1986
Eldoninto	ĈEE
Klarigoj	Literaturaĵoj de la plej reprezentaj verkistoj de sia tempo.
Formato	602 paĝoj, 21 cm



$xpath->query("//td/div/a/img/@src")



B. Scraping de données

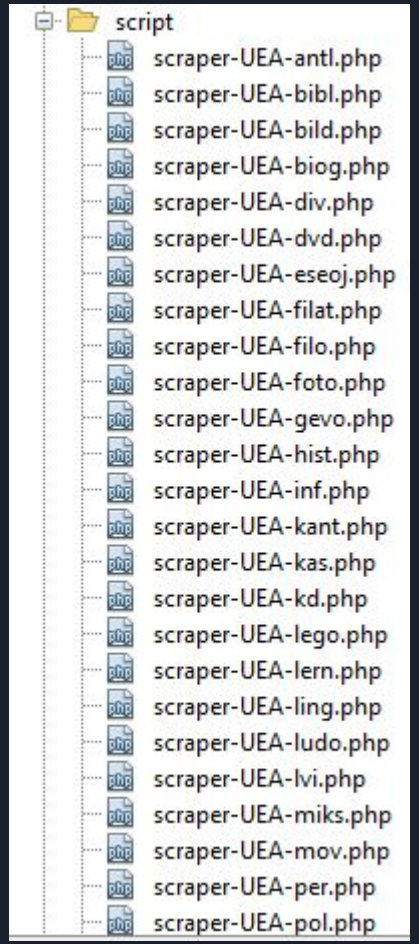
OpenLibrary

- **1100** oeuvres insérés

Katalogo

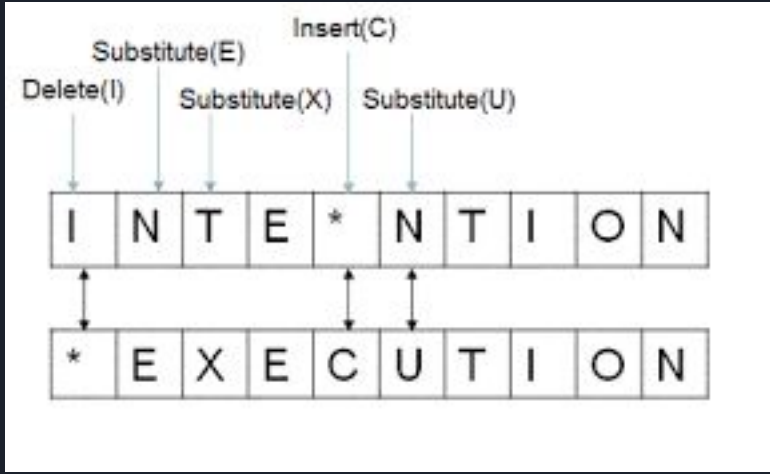
- **8200** oeuvres insérés

URL disponible pour chaque oeuvres



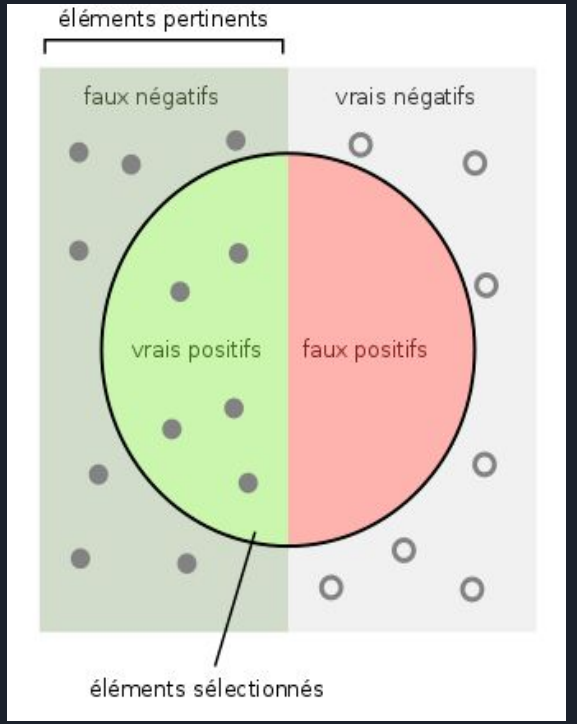
C. Déduplication

- Algorithme de Levenshtein : nombre d'opérations pour transformer une chaîne en une autre
- Seuil de détection
- Suppression doublon / Ajout dans une table



C. Déduplication

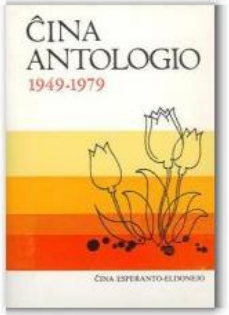
Nom du cas	Abréviation	Description
Vrai positif	VP	L'algorithme trouve **à raison** la paire d'oeuvres comme **appartenant** aux doublons
Faux positif	FP	L'algorithme trouve **à tort** la paire d'oeuvres comme **appartenant** aux doublons
Vrai négatif	VN	L'algorithme trouve **à raison** la paire d'oeuvres comme **n'appartenant pas** aux doublons
Faux négatif	FN	L'algorithme trouve **à tort** la paire d'oeuvres comme **n'appartenant pas** aux doublons



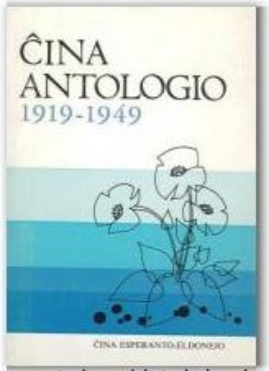
C. Déduplication

Exemple de FP

Titolo	Ĉina antologio (1949-1979)
Kategorio	Antologioj
Prezo	Nehavebla
Eldonloko, jaro	Pekino, 1989
Eldoninto	ĈEE
Klarigoj	Kvardeko da noveloj, eseoj kaj poemoj.
Kontribuantoj	Komp. Zhao Muying
ISBN/ISSN	7505200402
Formato	418 paĝoj, 22 cm



Titolo	Ĉina antologio (1919-1949)
Kategorio	Antologioj
Prezo	Nehavebla
Eldonloko, jaro	Pekino, 1986
Eldoninto	ĈEE
Klarigoj	Literaturaĵoj de la plej reprezentaj verkistoj de sia tempo.
Formato	602 paĝoj, 21 cm



C. Déduplication

- La précision compte la proportion d'items pertinents parmi les items sélectionnés
- Le rappel compte la proportion d'items pertinents sélectionnés parmi tous les items pertinents sélectionnables.

$$rappel = \frac{TP}{TP+FN} \qquad précision = \frac{TP}{TP+FP} \qquad F\,Mesure = \frac{précision * rappel}{précision+rappel}$$

	Algorithme 1	Algorithme 2	Algorithme 3	Algorithme 4
configuration (paramètres)	levenshtein($titre, $titre2, 1, 1, 1)	levenshtein($titre, $titre2, 1, 1, 1)	levenshtein($titre, $titre2, 3, 1, 1)	levenshtein($titre, $titre2, 3, 3, 3)
Pourcentage de différence entre les oeuvres (seuil)	< 20%	< 40%	< 40%	< 40%
précision	≃ 0.0069	≃ 0.0010	≃ 0.0018	≃ 0.0103
rappel	≃ 0.0833	≃ 0.0770	≃ 0.1	≃ 0.125
F-mesure	≃ 0.0064	≃ 0.0010	≃ 0.0017	≃ 0.0095

C. Développement d'un site web

- Visualisation des oeuvres en espéranto
- Oeuvres récentes sur l'accueil
- Multilingue (FR,EN,EPO)
- Responsive

PHP Version 8.1.1

A. Piste d'amélioration

- Nouvelles sources pour le scraping (VIAF, WorldCat,BNF, ...)
- Déduplication avec d'autres algorithmes (strcmp, similar_text)
- Filtre par catégorie (et autres)
- Demande d'ajout d'oeuvre

Démonstration