PROJET S.I : EURONEXT

Rendu numéro 2 : Applications relationnelles

Nicolas BONNIOL Mathilde COLAVITTI Amanda KRAJCO ChenYang GAO

Sommaire

Introduction

Nous rappelons alors le but de ce projet transversal d'IG4 : il s'agit de rechercher le nombre d'occurrences de plusieurs mots (en relation avec la Responsabilité Sociale des Entreprises) sur les sites internet des entreprises cotées à la bourse (pour les marchés parisiens et bruxellois).

Dans ce projet, une partie est dédiée à notre module « Applications Relationnelles » , et pour cette partie nous devions réfléchir sur la conception d'une base de données et sur la création de masques (interface pour l'utilisateur).

Nous avons donc utilisé :

- Win Design pour réaliser le Modèle conceptuel des données ainsi que le Modèle logique
- Oracle forms pour réaliser la base de données et les masques

Nous avons du être capable d'insérer dans notre base de données les caractéristiques de toutes les entreprises ainsi que le nombre d'occurrences des différents mots clés (liste préalablement établie).

Toutes les informations que nous utilisons ont été récoltées sur le site Euronext (site mondial de la bourse). Ces informations ont été exploitées grâce à un script PHP que nous expliquons au fil de notre rendu.

Toute notre base de données est consultable grâce aux identifiants suivants :

mathilde.colavitti

oracle

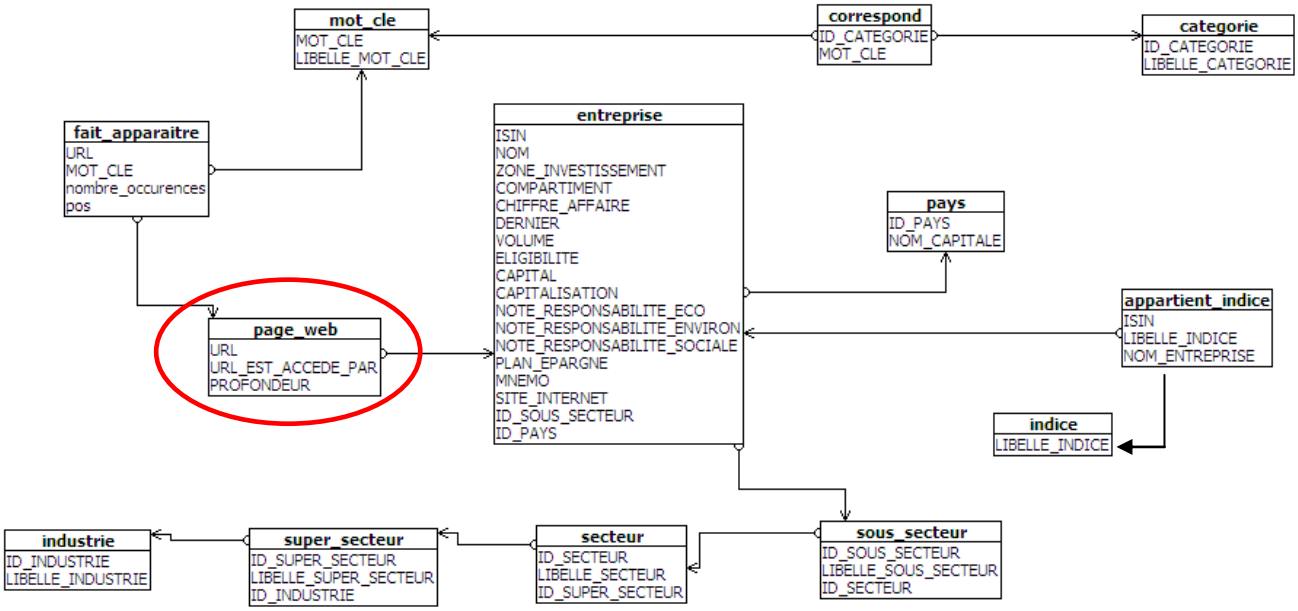
ORA10

Vous trouverez alors dans ce deuxième rendu les quelques modifications par rapport au premier rendu, les détails concernant le script PHP, et les étapes de notre élaboration des masques.

Modifications du MCD/MLR

Depuis le premier rendu, nous avons réétudié notre MCD et corrigé quelques erreurs, les plus importantes vous sont expliquées ci après :

- D'abord nous avons supprimé l'entité ***Capitalisation***, car on peut y avoir accès en utilisant une requête comme celle-ci : *select from entreprise where capitalisation between XX and XX*
- Ensuite, nous avons du modifier l'association ***Fait_Apparaître*** car avec la précédente nous ne pouvions ni connaître la phrase ni le nombre d'occurrences d'un mot. Cependant, un de nos souhaits les plus importants est de connaître le nombre de fois qu'un mot apparaît sur un site donné, et son contexte.
  Nous avons donc rajouté deux attributs : *nb_occurence* (int) et *phrase* (long)



- Dans l'entité **Page_Web**, nous avons modifié l'attribut *Isin* par un attribut *Url_est_accede_par* dans le but de connaître l'adresse de la page d'accueil liée.
  Exemple : *http://www.ast-groupe.fr/finance.php?varpage=liste-actualites.php&* est accédé par *http://www.ast-groupe.fr* .
  Nous avons aussi inséré un autre attribut qui est la *Profondeur*. Il s'agit là pour nous de connaître, grâce aux « / » présents dans une URL, la profondeur de la page.

- Enfin, nous avons décidé de créer une nouvelle table dans laquelle se trouvera l'indice boursier (après entretiens avec le demandeur nous avons compris que ce dernier était important pour l'étude statistique qui suivra).

Ainsi, une entité **Indice** avec l'attribut *Libelle_indice* et une association **Appartient_Indice** avec les attributs *Isin, Nom_entreprise et Libelle_indice* sont créées.

Une fois notre Modèle conceptuel de données et notre Modèle logique relationnel bien établis, nous avons pu commencer à remplir la base de données (grâce aux informations récupérées sur le site Euronext).

Explications techniques et récupération des données

D'abord, il est important de préciser que toute notre base de données est en Anglais.

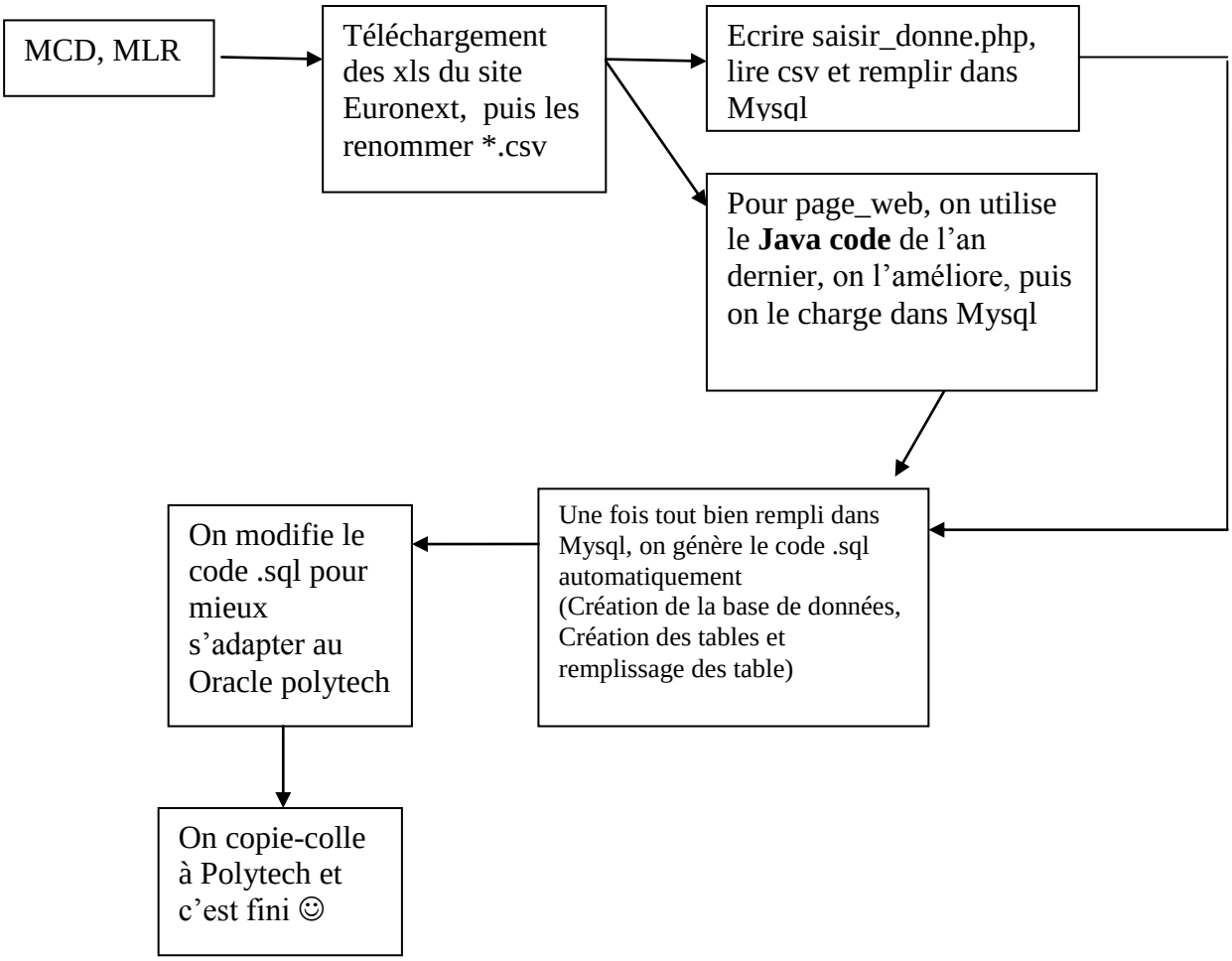
Pour remplir notre base de données à Polytech, nous avons utilisé : un script PHP, plusieurs fichiers Java et des requêtes PL/SQL.

Le script PHP a été créé par nos soins. Il sert à alimenter les tables comme SECTEUR, SOUS_SECTEUR, INDUSTRIE, … . Pour remplir ces tables, on utilise en fait un fichier csv.

Ensuite pour les autres tables comme APPARTIENT_INDICE, INDICE, CORRESPOND, CATEGORIE, …, le script PHP les a créé mais ensuite nous les avons remplies nous même à la main dans la base de données.

Pour ce qui concerne les fichiers Java, nous nous sommes inspirés des codes de l'année dernière que nous avons adaptés à notre base de données. Leur but est de parcourir les sites Internet et d'alimenter ensuite les tables suivantes : FAIT_APPARAITRE et PAGE_WEB.

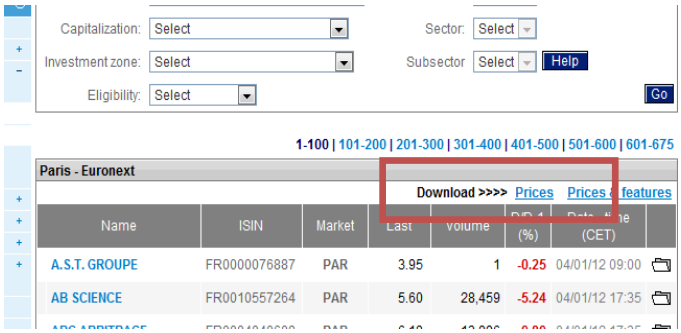
On aurait pu faire le script PHP directement sur la base de données Oracle, mais pour pouvoir travailler chez nous, nous avons préféré le faire de la manière suivante :



Comment ça marche ?

1. Récupérer les données

En premier lieu, on télécharge le fichier xls à partie du site Euronext en choisissant les deux marchés qui nous intéressent : Paris et Brussel. Ensuite on change le fichier xls en fichier csv.



2. Script PHP

Ensuite, dans le script PHP, nous avons écrit les fonctions suivantes :

- **function initialisation($conn)**
  Cette fonction a pour objectif d'initialiser la base de données. En fait, on crée un lien de connexion dans le code PHP pour pouvoir accéder à la base de données.
  Cette fonction sert aussi pour la création des tables, la suppression de tables et l'alimentation des tables (PAY).

- **function_insert_Categorie($conn)**
  Cette fonction sert à alimenter la table CATEGORIE, qui est ensuite remplie par nos soins, à la main.

- **function insert_Correspond($conn)**
  Cette fonction alimente la table CORRESPOND, qui est ensuite remplie par nos soins, à la main. Cette table représente le lien entre CATEGORIE et MOT_CLE.

- **function insert_Indice($conn)**
  Cette fonction alimente la table INDICE, que nous remplissons ensuite à la main (après avoir fait les recherches sur Internet). Les indices que nous avons rentrés sont les suivants : ASPI, AUCUN, DJSI, ESI et FTSI.

- **function insert_appartient_indice($conn)**
  Cette fonction alimente la table APPARTIENT_INDICE, que nous remplissons également à la main. Elle est le lien entre l'INDICE et l'ENTREPRISE.

- **function insert_mot_cles($chemain, $conn)**
  Cette fonction alimente la table MOT_CLE. Le contenu de cette table était défini par Mme BOURDON mais nous l'avons complété avec d'autres mots. Cette fonction lit un fichier txt et rempli la base de données.

- **function insert_isss($industrie_id, $industrie_nom, $super_secteur_id, $super_secteur_nom, $secteur_id, $secteur_nom, $sous_secteur_id, $sous_secteur_nom, $conn)**
  Cette fonction alimente les tables SOUS_SECTEUR, SECTEUR, SUPER_SECTEUR et INDUSTRIE, après avoir réussi à analyser le fichier csv (tiré du fichier xls d'internet).
  On appelle cette fonction dans la fonction *analyse_csv($chemain, $pay_id, $conn)*.

- **function analyse_csv($chemain, $pay_id, $conn)**
  C'est une fonction qui sert principalement à analyser le fichier csv pour pouvoir mettre à jour notre base de données et notamment les tables : ENTREPRISE, SOUS_SECTEUR, SECTEUR, SUPER_SECTEUR et INDUSTRIE.

On peut généraliser la démarche par la suite d'actions suivante :

Initialisation →lire le fichier dictionary.txt→remplissage de MOT_CLE→ remplissage de CATEGORIE→ remplissage de CORRESPOND→ remplissage d'INDICE→ remplissage de CORRESPOND_INDICE→analyser le fichier csv et remplir ENTREPRISE, SOUS_SECTEUR, SECTEUR, SUPER_SECTEUR et INDUSTRIE.

3. Code Java

Nous avons utilisé le code Java d'élèves de l'année dernière. Cependant nous l'avons modifié pour l'adapter à notre schéma (modèle conceptuel de données) et à notre base de données.

Nous avons quelques remarques à préciser :

- Lorsqu'on utilise le thread qui analyse les pages web, on met pour chaque thread 3 secondes. Cela suffit surement mais normalement c'est 5 secondes ou plus donc cela se peut que l'on ne puisse pas assurer pas la totalité des données. Ceci s'explique peut-être par le fait que quelques fois le site web n'est pas accessible à cause de problèmes Internet.

- Pour la profondeur, nous n'avons pas réussi à analyser au fur et à mesure de la recherche récursive des mots clés : ce que l'on a fait c'est uniquement mettre dans l'attribut 'profondeur' le nombre de '/' dans chaque URL. Du coup, ce n'est surement pas juste dans la base de données.

Par contre, nous avons amélioré le script Java pour mieux s'adapter à notre système et pouvoir bien analyser la position de chaque occurrence d'un mot clé (c'est-à-dire l'endroit ou un mot clé apparaît dans la page Web).
L'idée est la suivante : nous avons décidé d'attribuer à chaque Tags une note (que nous avons décidé nous même). Cela nous servira pour l'analyse statistique, car nous avons vu avec la demandeuse (Mme BOURDON) qu'attribuer des notes de communication à chaque entreprise était une idée pertinente.

La note maximum d'un tag est de 5 points, voici la répartition des notes en fonction des différents tags :

<h1> : il s'agit d'un gros titre, c'est donc une bonne note = 5 points
<h2> : il s'agit aussi gros titre mais moins important que <h1> = 4 points
<h3> : 3 points
<h4> : 2 points
<a> : il s'agit d'un lien qui renvoie à une autre page du site web = 1 point
<description > : il s'agit d'un mot enregistré pour qu'un moteur de recherche puisse afficher le site web lorsqu'on tape ce mot dans la barre de rechercher = 3 points
 : le mot apparaît dans un paragraphe ou une phrase normale. Le mot n'est donc pas trop visible pour les utilisateurs d'internet = 1 point
<img> : un image nommée par le mot clé = 2 points.
 : le mot apparaît dans une liste, c'est pratiquement équivalent à = 1 point
<p> : le mot apparaît dans un nouveau paragraphe, c'est équivalent à = 1 point
 : le mot apparaît mais avec une apparence particulière comme la couleur, le style (souligné, gras, italique, centré, …) = 2 points
<autre> = 0 point

Ainsi, pour attribuer la note de chaque entreprise, au lieu d'analyser la profondeur, nous analyserons l'apparence des différents tags pour chaque occurrence d'un mot.
Par exemple, une entreprise dont le site web est composé de 10pages(profondeur), pour le mot clé 'environnement' :
<h1>apparaît 5 fois, et apparaît 8 fois
On calcule la note Environnement de la façon suivante : 5*5+8*1=33
Et ainsi pour chaque entreprise, pour les trois notes que nous avons attribuées.

Construction des masques

L'objectif unique de la création de masques est de pouvoir donner aux utilisateurs la possibilité d'entrer en contact avec la base de données.
Nous avions réfléchi pour le premier rendu à différents masques. Cependant, nous avons décidé, avec le temps et en voyant les multiples contraintes qui se posaient, de les modifier.

Effectivement, nous avions pensé à réaliser un masque dans lequel on aurait pu modifier/rajouter un mot ou une catégorie de mot. Désormais ceci est pour nous impossible, dans le sens où, si on rajoute/modifie un mot ou une catégorie de mot, ce sera le fichier .csv que l'on obtient après mise en marche de notre script qu'il faudrait modifier.
Nous imposons donc l'interdiction d'insérer, de supprimer, et/ou de modifier n'importe quel mot clé ou n'importe qu'elle catégorie de mot.

En ce qui concerne les catégories de mots, nous nous sommes référencés à celles données par Mme BOURDON sur les documents, autrement dit les catégories suivantes :

- Catégorie responsabilité sociale des entreprises
- Catégorie environnement/écologie
- Catégorie chartes
- Catégorie certification
- Catégorie labels

Pour ce qui concerne les trois notes que nous avons décidé d'attribuer, nous avons changé les noms, il s'agit maintenant des trois suivantes :

- Note de **Responsabilité Sociale** (correspondant à la catégorie de mots responsabilité sociale des entreprises)
- Note d'**Environnement/Ecologie** (correspondant à la catégorie de mots environnement/écologie)
- Note d'**Engagement** (correspondant aux catégories de mots labels, certification, et chartes)

Maintenant que toutes les contraintes ont été gérées, l'utilisateur, grâce aux masques, pourra faire différentes choses :

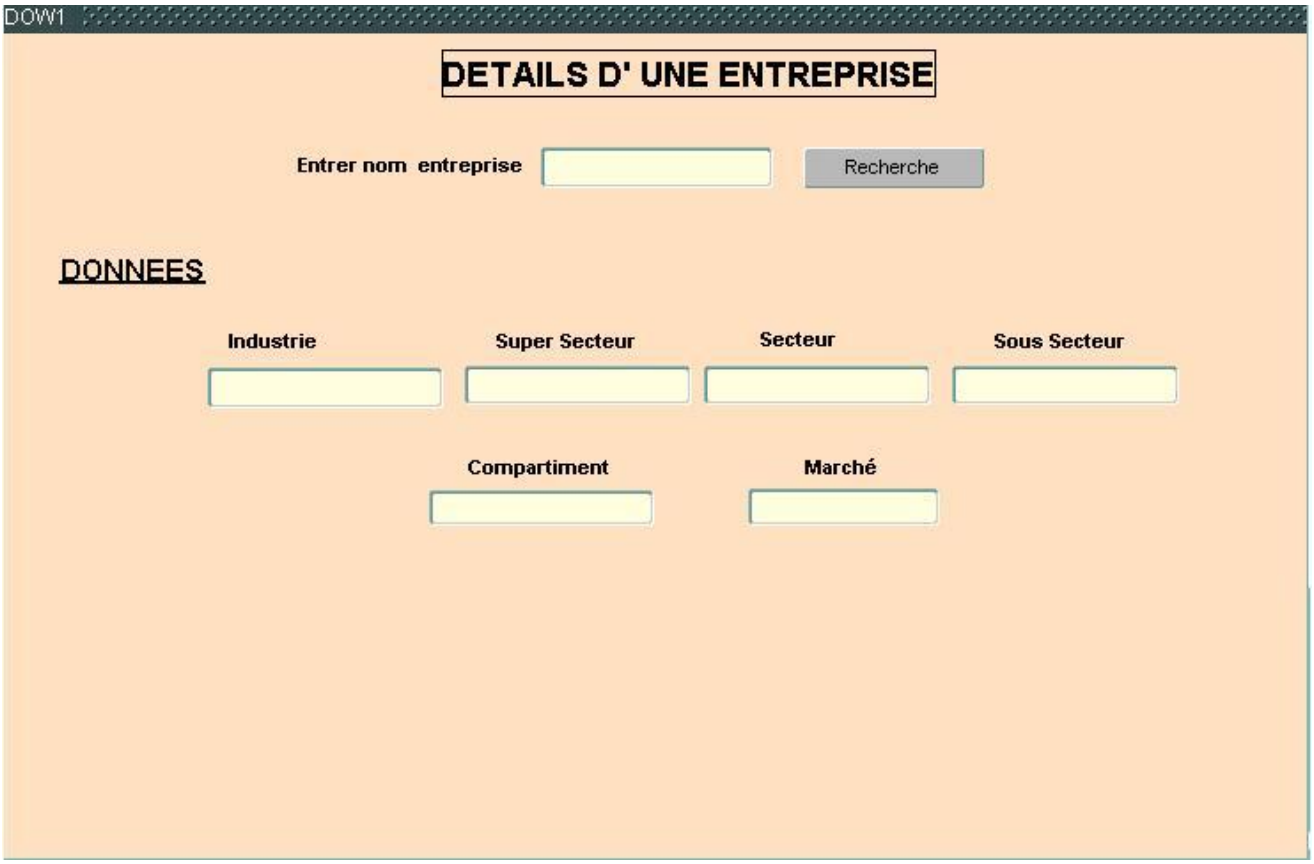
- Consulter les détails d'une entreprise : son industrie, son super secteur, son secteur, ….
- Consulter, à partir d'une industrie/d'un super secteur/d'un secteur/ d'un sous secteur ou d'une entreprise, la liste des mots clés présents et leur nombre d'occurrences.
- Consulter la liste des entreprises/ industrie/ super secteur/ secteur/ sous secteur pour lesquels un mot donné apparaît
- Consulter les indices boursiers de chaque entreprise, et les ajouter/supprimer/modifier.

1. Premier masque : consultation des détails d'une entreprise

Avec ce premier masque l'utilisateur pourra consulter les informations concernant une entreprise qu'il donnera.

Il devra rentrer le nom de l'entreprise (et non son Isin) correctement. Si l'entreprise rentrée n'existe pas, un message d'erreur sera affiché.

L'utilisateur pourra donc avoir accès aux détails suivant de l'entreprise :
Marché, compartiment, secteur, sous secteur, super secteur, site web, capitalisation, et autres.



2. Deuxième masque : mots clés et occurrences par industrie/super secteur/secteur/sous secteur/entreprises

Ce second masque aura la fonction suivante : afficher la liste des mots clés pour une industrie, un super secteur, un secteur ou une entreprise. En plus d'afficher la liste des mots clés, le masque affichera à la suite de chaque mot son nombre d'occurrences dans l'industrie, super secteur, secteur, sous secteur ou entreprise.

L'utilisateur devra rentrer, selon son souhait, exactement le ou les libellés de chaque industrie/super secteur/secteur/sous secteur. Pour l'entreprise, il devra saisir son nom exact aussi.

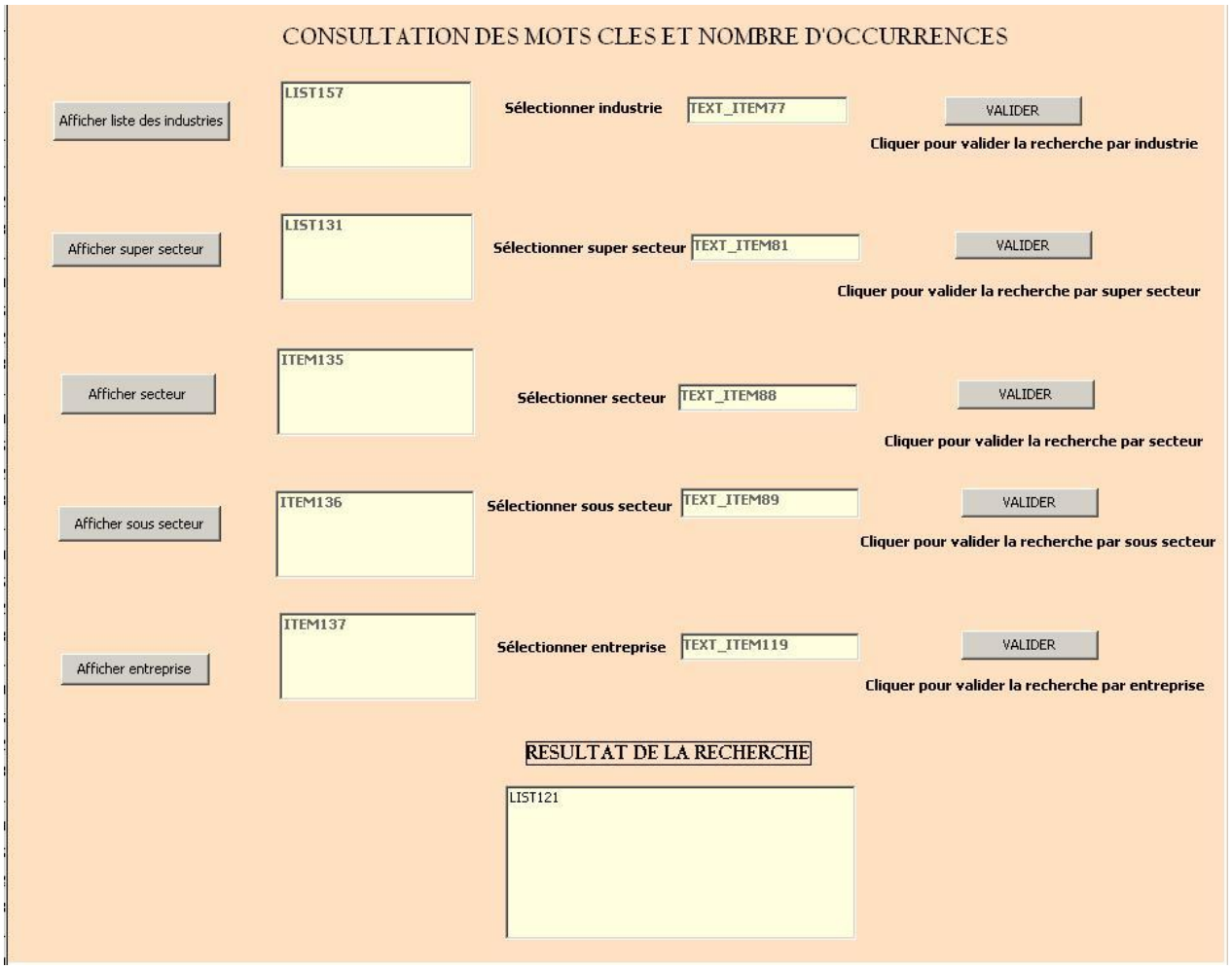
Les champs doivent être remplis au fur et à mesure : par exemple l'utilisateur ne pourra afficher les mots clés d'un secteur uniquement après avoir rempli au préalable l'industrie, le super secteur et enfin le secteur.

Le clique sur le bouton valider (du secteur) déclenchera la requête PL/SQL et affichera donc les mots clés et leurs occurrences.

Et ainsi de suite, selon le niveau pour lequel il souhaite consulter les mots clés.

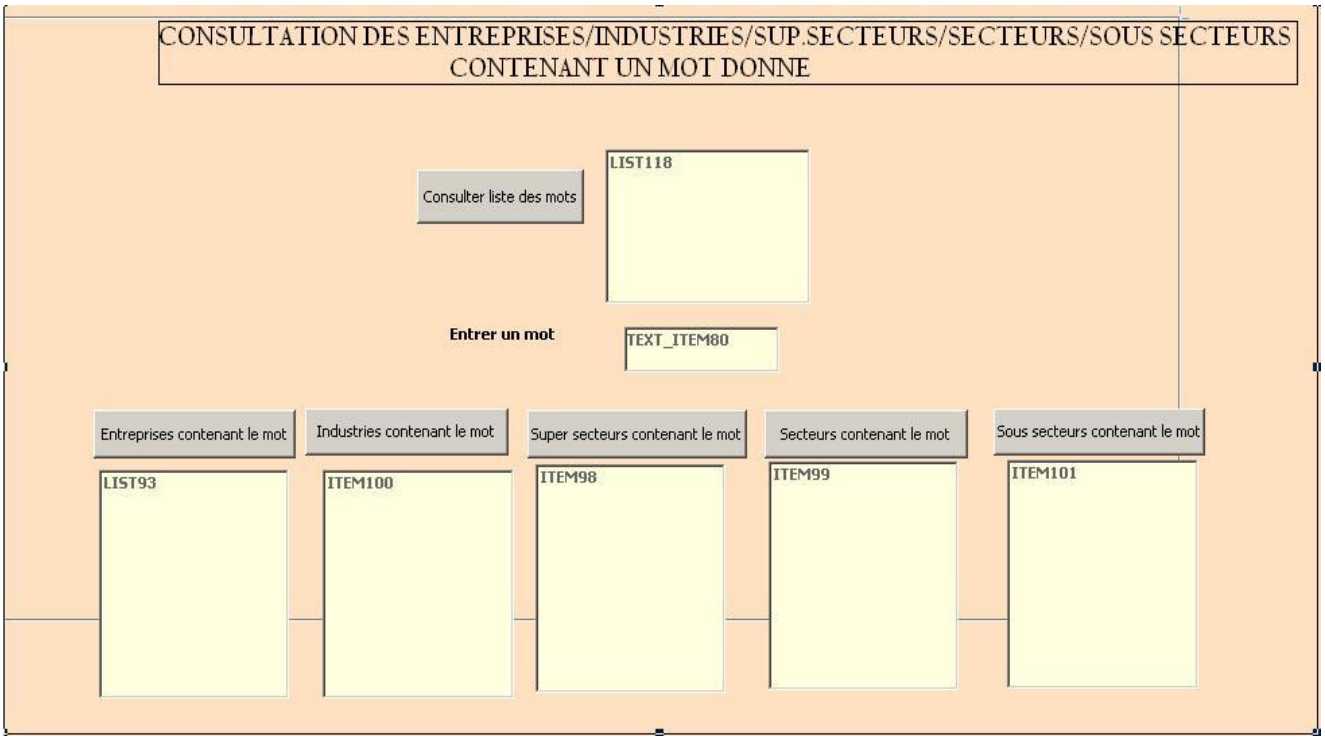
Nous avons bien pris en compte que l'utilisateur ne peut pas connaître tous les noms exacts d'industries/de super secteurs/de secteurs/ de sous secteurs/ d'entreprises. C'est pour cela que nous avons décidé d'insérer un catalogue des noms : il faudra pour cela à l'utilisateur, afficher les industries puis en choisir une et la réécrire dans la case Sélection d'industrie, ensuite afficher les super secteurs puis en choisir un, ensuite afficher les secteurs puis en choisir un et ainsi de suite jusqu'à entreprise.

Attention, si l'utilisateur souhaite faire une nouvelle recherche il sera obligé de relancer l'application car sinon l'affichage des mots clés sera en double.


CONSULTATION DES MOTS CLES ET NOMBRE D'OCCURRENCES
Afficher liste des industries
LIST157
Sélectionner industrie
TEXT_ITEM77
VALIDER
Cliquer pour valider la recherche par industrie
Afficher super secteur
LIST131
Sélectionner super secteur
TEXT_ITEM81
VALIDER
Cliquer pour valider la recherche par super secteur
Afficher secteur
ITEM135
Sélectionner secteur
TEXT_ITEM88
VALIDER
Cliquer pour valider la recherche par secteur
Afficher sous secteur
ITEM136
Sélectionner sous secteur
TEXT_ITEM89
VALIDER
Cliquer pour valider la recherche par sous secteur
Afficher entreprise
ITEM137
Sélectionner entreprise
TEXT_ITEM119
VALIDER
Cliquer pour valider la recherche par entreprise
RESULTAT DE LA RECHERCHE
LIST121

3. Troisième masque : afficher entreprises ou industries ou super secteurs ou secteurs ou sous secteurs en fonction d'un mot

Dans ce troisième masque, l'utilisateur choisit un mot (qu'il trouvera dans un catalogue des mots), et le masque retournera alors soit la liste des entreprises et/ou la liste des industries et/ou la liste des super secteurs et/ou la liste des secteurs et/ou la liste des sous secteurs pour lesquels le mot sélectionné apparaît.

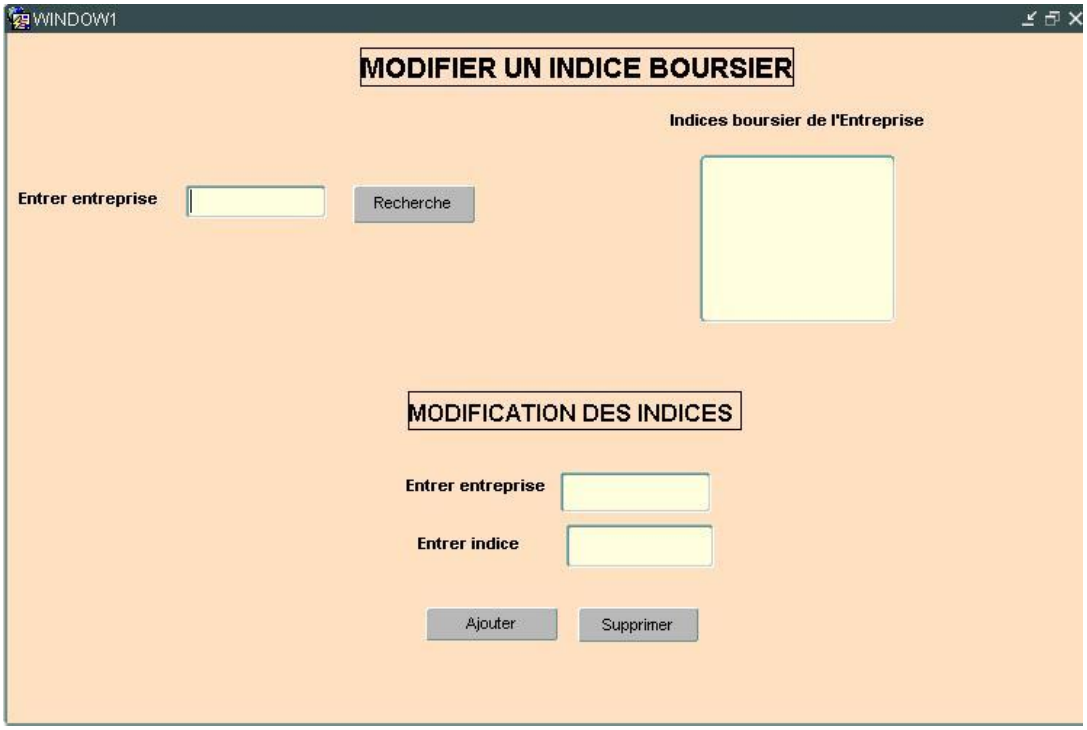


4. Quatrième masque : indices boursiers

Après avoir effectué des recherches sur internet, nous avons rentré manuellement tous les indices boursiers de chaque entreprise. Une entreprise n'appartient pas forcément à un indice, mais elle peut aussi en avoir plusieurs.
Le problème est le suivant : la liste des entreprises appartenant à un indice varie très souvent. Nous avons donc voulu donner la possibilité à l'utilisateur d'ajouter un indice pour une entreprise, ou encore d'en supprimer un ou plusieurs.

Attention, il sera possible de créer un indice boursier, et non pas seulement d'ajouter un indice parmi ceux que nous avons rentré dans la base de données (c'est-à-dire les 5 suivants : FTSE – ESI - ASPI – DSGI – AUCUN).

En ce qui concerne la fonctionnalité de ce masque : il suffit pour l'utilisateur de rentrer le nom d'une entreprise. Les indices boursiers auxquels elle appartient s'afficheront donc, et il sera alors possible pour l'utilisateur d'ajouter/supprimer un ou plusieurs indices.



5. Problèmes rencontrés et améliorations

Nous avions, à l'origine, souhaité réaliser des listes se mettant à jour dynamiquement. Cependant, nous n'avons pas réussi à implémenter à cause du peu de connaissances du logiciel Oracle Forms.

De plus, un des problèmes les plus importants est que nous n'avons pas gérer la différence entre majuscules et minuscules.

Aussi, nous avons effectué des requêtes dans la base de données, avec SQLplus, qui marchaient mais lorsque nous passions sur Oracle Forms il nous était impossible d'obtenir le même résultat avec l'assistant bloc de données. C'est pour cela que nous avons perdu du temps et donc pas réalisé tout ce que nous souhaitions.

Effectivement, nous avons réfléchi à la possibilité d'introduire une notion que nous pensions importante :

Il s'agit de la densité d'une page. Nous souhaitions compter le nombre de mots d'une page pour savoir si le nombre d'occurrences était ou non proportionnel à la taille de la page. Cela nous aurait permis de réaliser des statistiques bien plus poussées.