Master Intégration de Compétences

Enoncé du projet (première version)

Vincent Berry - Pierre Pompidor

Le but du projet est de réaliser un petit moteur de recherche sur un ou plusieurs documents formatés en XML.

Formatage XML des données :

XML (eXtensible Mark-up language) est un format de données (notamment pour des données textuelles) très apprécié, car relativement lisible et structuré pour l'humain, il permet d'opérer des parcours d'arbres très efficaces en informatique. Le choix des balises est libre (à moins de se référer à des taxonomies publiques de balises), mais doit présenter la plus grande sémantique possible.

```
<?xml version="1.0" encoding="UTF-8" ?>
<baliseX>
   <baliseY>
      ...
   </baliseY>
   <baliseY>
      ...
   </baliseY>
   <baliseZ>
      ...
   </baliseZ>
   ...
</baliseX>
```

Toutes les balises peuvent être également accompagnées d'attributs...

Nous vous proposerons ultérieurement des fichiers XML pouvons avoir un intérêt dans le cadre de votre première compétence (enfin nous l'espérons), mais rien ne vous empêche d'en chercher ou d'en créer par vous-même...

Dans ce qui suit, nous appellerons paragraphe un bloc de texte délimité au plus proche par deux mêmes balises (ouvrante et fermante).

Fonctionnalités (par étapes) obligatoires du projet (bref le strict minimum) :

Recherche dans un seul document :

- un mot est saisi dans une zone de texte → les paragraphes dans lesquels il se trouvent sont affichés.
- plusieurs mots sont saisis dans une zone de texte → les paragraphes dans lesquels ils se trouvent sont affichés;
- une balise est choisie dans la liste déroulante présentant les balises du document
  → les paragraphes circonscrits par cette balise sont affichés;
- un ou plusieurs mots sont saisis dans la zone de texte et une balise est choisie dans la liste déroulante
  → les paragraphes contenant ce ou ces mots, et circonscrits par cette balise, sont affichés;

L'interface graphique permettant la spécification des recherches doit s'afficher dans le navigateur, les scripts invoqués sur le serveur doivent être écrits en Perl (vive les expression srégulières)

Fonctionnalités facultatives (et non ordonnées et non exhaustives...) du projet :

- les fonctionnalités précédentes peuvent s'appliquer sur plusieurs documents XML, les noms de ces documents doivent alors être spécifiés dans un fichier de configuration (et non codés directement dans un script) ;
- la recherche peut être organisée suivant la sélection de plusieurs balises (et donc en opérant plusieurs filtres) ;
- les mots saisis peuvent être lemmatisés ou radicalisés pour lancer une recherche plus générique;
- des noms et des valeurs attributs peuvent être pris en compte dasn les recherches.