TP3 : manipulation XML

UCBL - Base de données pour la bioinformatique - 2024 / 2025

Objectif du TP : écrire des requêtes XPath et XQuery avec le SGBD BaseX

1 Préparation de l'environnement BaseX

Dans ce TP, nous utiliserons le SGBD BaseX[1], qui a l'avantage d'être automatiquement configuré, open-source, multi-plateformes, et fourni avec une interface graphique conviviale.

- **Téléchargement de BaseX** pour votre système : https://basex.org/download/
- **Démarrage de BaseX**. Lancez l'exécutable BaseX (double-clic ou `java -jar BaseX.jar`).

2 Création du jeu de données

Dans cette partie, nous créons le jeu de données, qui concerne la description d'un génome en XML[2]. BaseX considère un fichier XML comme une base de données.

- **Enregistrez le document XML** : https://perso.liris.cnrs.fr/fabien.duchateau/ens/BDBIO/tp/tp3.xml
- **Créez une nouvelle BD** en choisissant le menu (`Database > New…`). Dans la fenêtre de dialogue, sélectionner le fichier XML que vous venez d'enregistrer. Dans l'onglet `Parsing`, cochez les options `use internal parser` et `strip namespaces`. Cliquez sur OK et BaseX crée alors une base de données avec le fichier fourni et vous y connecte. Vous remarquerez que BaseX propose différentes visualisations des données, notamment celles d'arbre et de map qui permettent de naviguer facilement.

3 Exécution de requêtes XPath / XQuery

Dans cette dernière étape, il faut écrire les requêtes suivantes en XPath ou XQuery[3]. Vous pouvez écrire vos requêtes dans un fichier texte ou directement dans l'éditeur de BaseX. Dans les deux cas, n'oubliez pas de sauvegarder régulièrement. Pour exécuter une requête, il faut la mettre dans l'éditeur et cliquer sur le bouton *Run query*.

Requêtes à traduire en XPath ou XQuery :

1. L'ensemble du document
2. Les (balises) protéines (79 éléments résultat)
3. Les textes (sans balises) des noms recommandés de protéines (122 ou 83 ou 124 éléments résultat)
4. Les références vers d'autres bases (balise *dbReference*) pour les organismes (79 éléments résultat)
5. Les citations qui n'ont pas été publiées dans un article de journal (1 élément résultat)
6. Le titre des articles publiés après 2008 (47 éléments résultat)
7. Le nombre de citations (résultat = 357)
8. Le nombre d'éléments supportés par un attribut *evidence* (résultat = 1131)

[1] Site officiel https://basex.org/, documentation sur https://docs.basex.org/wiki/Main_Page

[2] Source Uniprot [10377 reviewed], https://www.uniprot.org/uniprotkb?query=%28taxonomy_id%3A10377%29

[3] Si aucune requête ne fonctionne (0 résultat, y compris des requêtes comme `/`), alors il faut supprimer la BD et la recréer en cochant les options *Use internal XML parser* et *Strip namespaces* de l'onglet *Parsing*

9. Le titre des références dont les mots-clés (balise *scope*) contiennent *CRYSTALLOGRAPHY* (21 éléments résultat)
10. Les liens HTML vers les références (balise *dbReference*) de type *NCBI Taxonomy* (2 résultats). La référence sert à la fois à construire l'URL et comme ancre du lien. Exemple du format de sortie :
    ```
    <a href="https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=12345">12345</a>
    ```
11. La liste des références Uniprot (*accessions*) pour chaque protéine (79 résultats). Le nom de protéine sera stockée comme attribut d'une balise `<accessions>`. Cette balise contiendra des balises `<refs>` dont le texte sera la référence Uniprot. Exemple de format de sortie :
    ```
    <accessions protein="abcdef">
         <ref>p12345</ref>
         <ref>q67890</ref>
    </accession>
    ```
12. Classement des personnes qui signent un article en première position (186 résultats). Le nom de la personne apparait comme attribut d'une balise *nbFirstAuthor*, et le texte de la balise correspond au nombre de signatures en tant que première auteur. Les balises seront triées par ordre décroissant d'articles signés en première position. Exemple de format de sortie :
    ```
    <nbFirstAuthor name="Jane Doe">42</nbFirstAuthor>
    <nbFirstAuthor name="John Doe">24</nbFirstAuthor>
    ```