Ecole d'ingénieurs et d'architectes de Fribourg
Hochschule für Technik und Architektur Freiburg
University of Applied Sciences of Fribourg

Travail pratique eXist
Annexes – Paramétrage des index

(Documentation eXist: search index

search lucene)

1 Introduction

The current version of eXist by default includes the following types of indexes:

1. **Structural Indexes** : These index the nodal structure, elements (tags) and attributes, of the documents in a collection.
2. **Range Indexes** : These map specific text nodes and attributes of the documents in a collection to typed values.
3. **Old Legacy Full Text Indexes** These map specific text nodes and attributes of the documents in a collection to text tokens.
4. **New Full Text Indexes (eXist 1.4):** new full text indexing module. Features faster and customizable full text indexing by transparently integrating Lucene into the XQuery engine. *Prefer this index over the old builtin implementation.*
5. **NGram Indexes :** These map specific text nodes and attributes of the documents in a collection to splitted tokens of n-characters (where n = 3 by default). Very efficient for exact substring searches and for queries on scripts (mostly non-european ones) which can not be easily split into whitespace separated tokens and are thus a bad match for the full text index.
6. **Spatial Indexes *(Experimental)*:** These map elements of the documents in a collection containing georeferenced geometries to dedicated data structures that allow efficient spatial queries.

2 Configuration des index dans eXist

Les index sous eXist sont paramétrés avec des fichiers XML portant l'extension « .xconf ».

Ce XML est structuré de la manière suivante :

```
<collection xmlns="http://exist-db.org/collection-config/1.0">
   <index xmlns:atom= "http://www.w3.org/2005/Atom"
      xmlns:html="http://www.w3.org/1999/xhtml"
      xmlns:wiki="http://exist-db.org/xquery/wiki">
      <!-- Disable the standard full text index -->
      <fulltext default="none" attributes="no"/>
      <!-- Lucene index is configured below -->
      <lucene>
          <analyzer class="org.apache.lucene.analysis.standard.StandardAnalyzer" />
             <analyzer id="ws" class="org.apache.lucene.analysis.WhitespaceAnalyzer" />
           <text qname="FILM" />
      </lucene>
   </index>
</collection>
```

La balise root est « collection ». Elle contient une balise « index » permettant de décrire les index qui peuvent être de plusieurs types, chacun de ceux-ci ayant une balise propre. Pour un index de type « Lucene », la balise correspondante est « Lucene ».

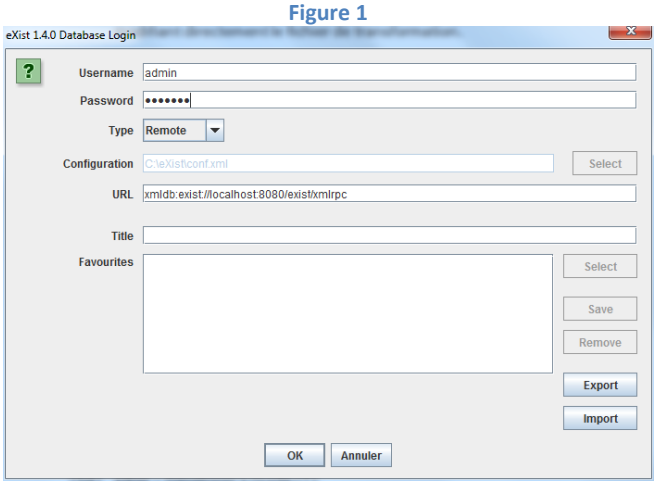
Dans l'exemple ci-dessus, l'index « Lucene » se porte sur la balise « FILM » des fichiers XML.

L'indexation en « texte intégral » (Lucene) est utilisée pour faire des recherches par mot clé sur de grandes collections très rapidement. Quand un fichier est ajouté à la collection, eXist l'indexe.

3 Paramétrage des index

Une fois un fichier d'index créé, il faut l'insérer dans la base de données eXist. Pour cela, vous pouvez utiliser l'outil « client.bat » que vous trouvez sous « <install_path>/eXist/bin » ou via le « eXist client Shell » du menu démarrer.

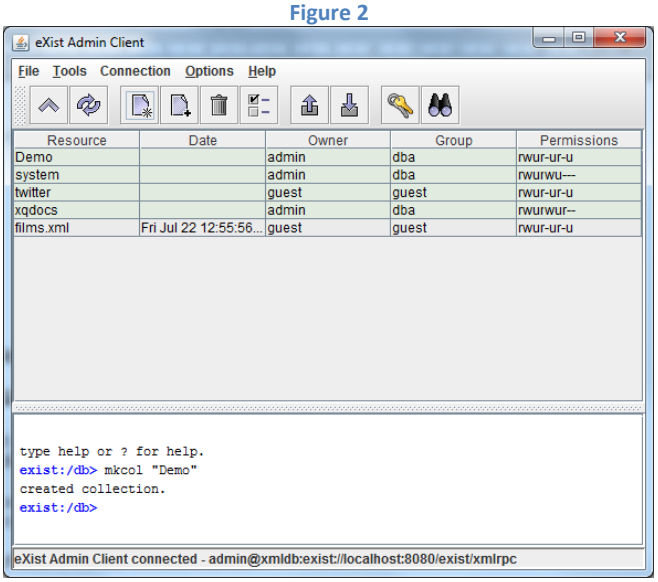
Figure 1



La première chose à faire est de se connecter à la base de données « eXist » (cf. Figure 1).

Pour cela, vous devez entrer votre mot de passe d'administration d'eXist et cliquer sur le bouton « OK ».

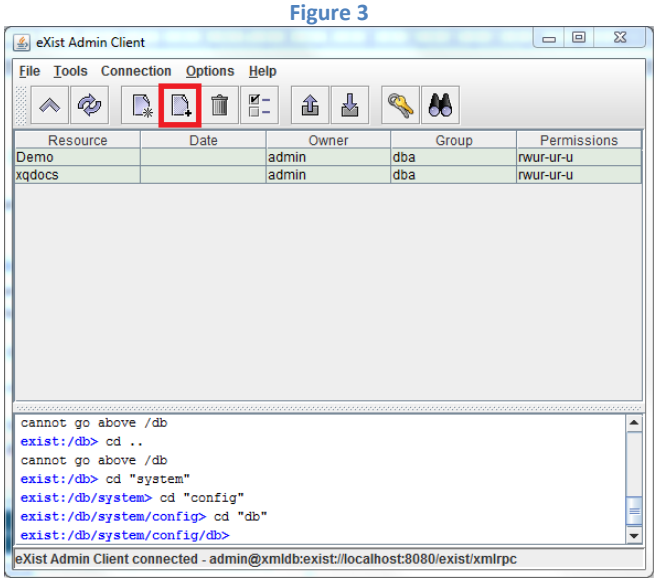
Figure 2



Sur la page qui s'ouvre (cf. Figure 2), vous pouvez voir le contenu de la base de données. Ici, la base contient 4 collections (Demo, system, twitter et xqdocs) et 1 fichier de données (films.xml)

L'insertion de l'index doit se faire dans un endroit particulier**. Il faut aller sous la collection « system / config / db »** en double cliquant sur le nom des collections.

Figure 3

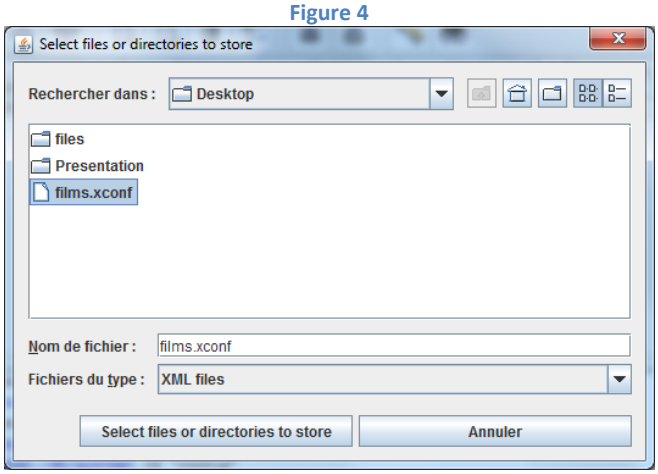


Le contenu de cette collection (cf. Figure 3) doit suivre la même architecture que la collection de base (cf. Figure 2). Comme le montre ces deux Figure, il y a dans les deux cas les collections « Demo » et « xpdocs ».

Pour ajouter l'index à la racine de la base de données, vous devez cliquer sur le bouton « store one or more files to the database » (dans le cadre rouge, cf.

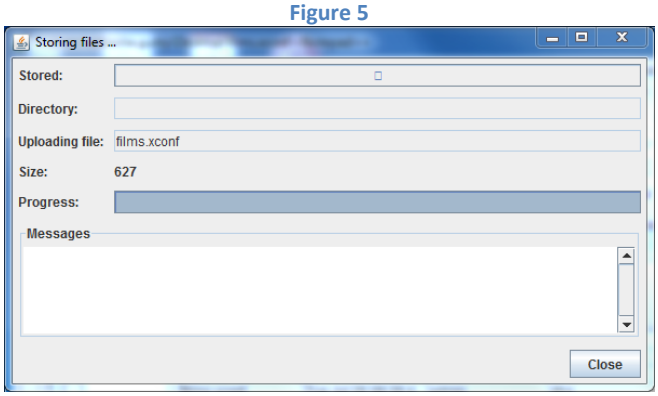
Figure 3). Si vous voulez ajouter l'index à une collection déterminée, vous devez préalablement aller dans la collection, par exemple « Demo ».

Figure 4



Vous devez ensuite sélectionner le fichier à charger dans la fenêtre qui s'ouvre (cf. Figure 4), puis cliquer sur le bouton « Select files or directories to store ».

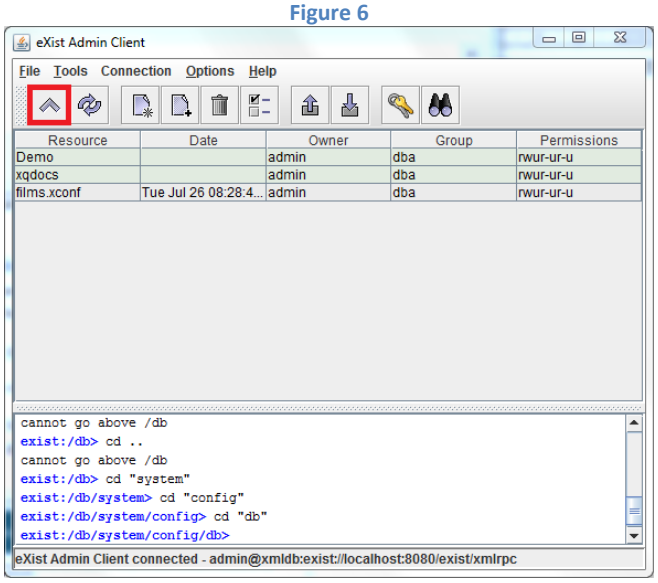
Figure 5



Remarque : au lieu de charger le fichier vous pouvez également le créer directement dans la fenêtre client (cf. Doc annexe installation eXist « 8.1 Création de collections et chargements de fichiers »)

Une fenêtre montrant le chargement du fichier s'ouvre alors (cf. Figure 5). Une fois le chargement complété, vous devez appuyer sur le bouton « Close ».

Figure 6

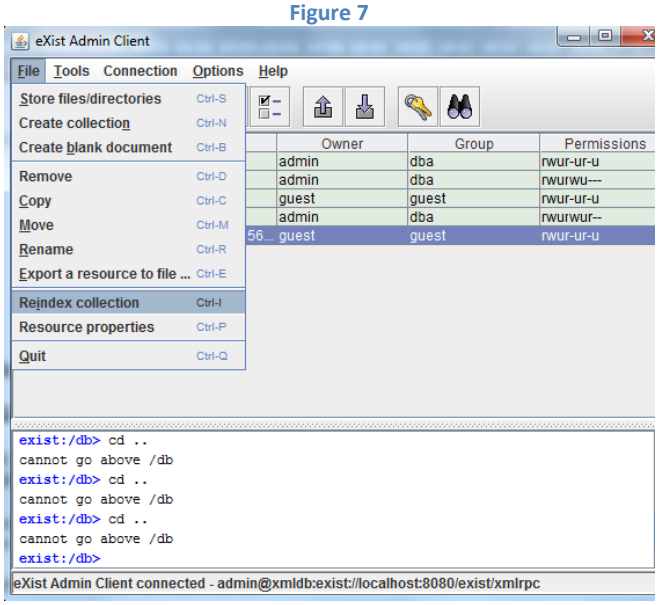


Vous pouvez ensuite voir que le fichier a bien été chargé (cf. Figure 6).

L'index est maintenant actif pour toutes les nouvelles données que vous insérerez dans la collection de la base de données. **Par contre, pour les données déjà insérées, l'index n'est pas actif.** Il faut manuellement dire à eXist d'indexer le contenu déjà existant.

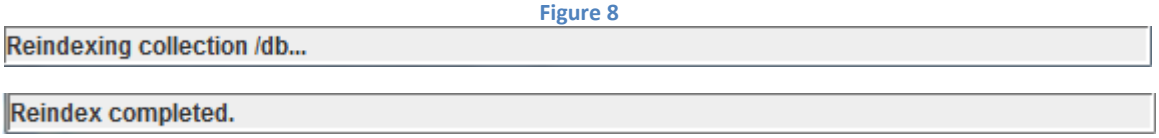
Pour cela, vous devez retourner à la racine de la base de données en **cliquant 3 fois sur le bouton « Go to parent collection »** (dans le cadre rouge, cf. Figure 6).

Figure 7



Sur cette page (cf. Figure 7), vous devez cliquer sur le menu « File », puis « Reindex collection » pour indexer le contenu déjà inséré dans la base de données. Un message de confirmation va apparaître. Cliquez sur « Oui »

Figure 8



Dans la barre de statut (cf. Figure 8), vous pouvez voir l'avancement de l'indexation. Une fois le message « Reindex completed » apparu, les données sont indexées.

4 Pour modifier le paramétrage des index

Pour modifier la façon dont une collection est indexée, un administrateur doit éditer le fichier de configuration associé à la collection qui se trouve dans les fichiers du système d'eXist.

Le fichier de configuration indique au système les règles à suivre lors de l'indexation, quelles sont les parties à indexer en « texte intégral », les balises à ignorer et même l'importance de certaines balises par rapport à d'autres pour la fonction de score.

Ne pas oublier de réindexer.