XSearcher : Un Système de Stockage et d'Interrogation de documents structurés en XML

Bessai Mechmache F.Zohra (*), Alimazighi Zaia (**), Slimani Farid (*), Belmansour A. Tidjani (*)

*(*) CERIST (**)Université des Sciences et de la Technologie Houari Boumediene (USTHB)*

E-Mail: {zbessai, slimani, tidjani}@cerist.dz , Alimazighi@wissal.dz

1- Introduction

Nous vivons dans un monde où l'information est omniprésente, que ce soit dans les journaux, dans les magazines, à la télévision ou sur Internet, l'information est disponible dans une multitude de formats tous plus accessibles les uns que les autres.

Augmenter la quantité d'information qui déferle sur tout un chacun peut sembler une amélioration de la qualité de vie, mais au-delà d'une certaine limite, des outils s'avèrent nécessaires pour pouvoir comprendre cette quantité d'information, l'analyser, en tirer profit et même en rejeter lorsque cela s'avère nécessaire. Ces outils sont des systèmes de recherche d'information dont le rôle est de fournir les techniques et les modalités permettant de sélectionner l'information pertinente répondant à des besoins d'information.

En ce qui concerne les systèmes de recherche d'information, on en distingue deux classes principales : les systèmes de recherche d'information traditionnels et les systèmes de recherche d'information structurée.

Pour les systèmes de recherche d'information traditionnels, c'est le document dans sa totalité qui est considéré comme entité atomique servant de base à la recherche et c'est donc le document entier qui est présenté à l'utilisateur comme réponse à sa requête. Cependant, les documents fournissent souvent une structure définie par l'auteur (par exemple, un document peut avoir plusieurs sections, chacune avec plusieurs sous-sections et ainsi de suite) et ce sont les éléments de cette structure qui sont considérés comme entités atomiques servant de base à la recherche dans les systèmes de recherche d'information structurée et par conséquent, seules les parties du document appropriées à un besoin d'information spécifique seront retournées à l'utilisateur [9] [12].

Ainsi, l'association des formats de données structurées (tels que XML) et des systèmes de recherche d'information structurée augmente la pertinence des résultats de recherche et fournit à l'utilisateur les informations qui répondent au mieux à ses besoins [10][11][13].

Le présent papier présente la conception et la réalisation d'un système de stockage et d'interrogation de documents au format XML basé sur le langage XQuery (ou XML Query) et ce, à travers une interface graphique et ergonomique. L'intérêt du présent travail est de montrer que grâce à la structuration logique du document, l'utilisateur va pouvoir consulter des documents plus pertinents pour sa recherche par :

- l'accès plus efficace aux documents susceptibles de l'intéresser,
- la recherche et la sélection des parties les plus pertinentes dans le document.

Pour cela, nous commencerons par un bref aperçu sur le langage XQuery. Nous traiterons, par la suite, la conception et la réalisation de notre système de stockage et d'interrogation de documents au format XML dénommé « XSearcher », en mettant en évidence la méthodologie de la réalisation, de l'expérimentation, ainsi que les résultats de ces expérimentations. Enfin, nous terminerons par une conclusion.

2- Le langage Xquery

Le langage Xquery [3] [4] [7] [8] est doté de mécanismes puissants associés à une syntaxe simple permettant une recherche fine et précise de l'information. Il permet de faire des interrogations aussi bien sur le contenu des données XML [1] que sur leur structure ou encore sur les deux en même temps. C'est un langage qui prend comme données d'entrée des structures XML et fournit aussi comme résultats des structures XML.

Il existe aujourd'hui des systèmes permettant l'interrogation des données XML [2] [6] en utilisant le langage XQuery qui sont, pour la plupart, disponibles sur le Web. Cependant, ces systèmes obligent l'utilisateur à saisir lui-même ses requêtes et donc à apprendre le langage XQuery. Cette contrainte restreint l'utilisation de ces systèmes à une certaine classe d'utilisateurs. Notre travail a un double objectif, surpasser cette contrainte en proposant un Système capable de fournir la puissance de XQuery en matière d'interrogation à travers une interface graphique conviviale et permettre aux utilisateurs une recherche d'information simple, fine et précise, soit par contenu (recherche plein texte), ou bien par structure (recherche de certaines parties du document).

3- Le Système Xsearcher

XSearcher est un système de stockage et d'interrogation de documents XML, basé sur le langage XQuery. Ce Système représente le fruit d'une pensée et ce distingue par

rapport à d'autres systèmes de recherche d'information par le fait que la recherche ne se fait pas d'une manière aléatoire mais d'une manière fine en permettant une recherche sur la structure du document (rechercher un élément d'après sa position dans un document) ou bien une recherche par contenus dans tout ou une partie du document.

Le système XSearcher se compose de trois modes d'interrogation et de deux modes d'importation des documents.
Les trois modes d'interrogation sont :

- Interrogation par contenus
- Interrogation par contraintes structurelles simples
- Interrogation par contraintes structurelles avancées

Les deux modes d'importation sont :

- Importation d'un document XML
- Importation d'un dossier contenant des documents XML

3.1- Architecture du Système XSearcher

XSearcher se compose principalement de quatre modules:

- **Le module d'interrogation** : qui se charge de générer automatiquement des requêtes XQuery à partir de la spécification des besoins de l'utilisateur, et de les exécuter.
- **Le module d'importation** : qui se charge de l'importation des documents dans la base de documents.
- **Le module d'enregistrement** : qui se charge de sauvegarder les résultats des interrogations en tant que nouveaux documents XML. Pour cela, l'utilisateur n'aura qu'à indiquer l'emplacement du nouveau document XML contenant les résultats et le module d'enregistrement se chargera de créer le document et d'y écrire les informations.
- **Le module d'affichage des résultats** : qui se charge d'afficher les résultats, dans un navigateur Internet, soit sous forme XML (avec le balisage), soit en appliquant une feuille de style CSS (Cascading StyleSheet) afin de procurer un affichage plus agréable en masquant le balisage des éléments. Nous avons opté pour ce type de feuille de style car CSS agit sur des pages HTML qui sont reconnues par tous les navigateurs web.

La figure ci-dessous illustre l'architecture du système par le diagramme de cas d'utilisations d'UML.

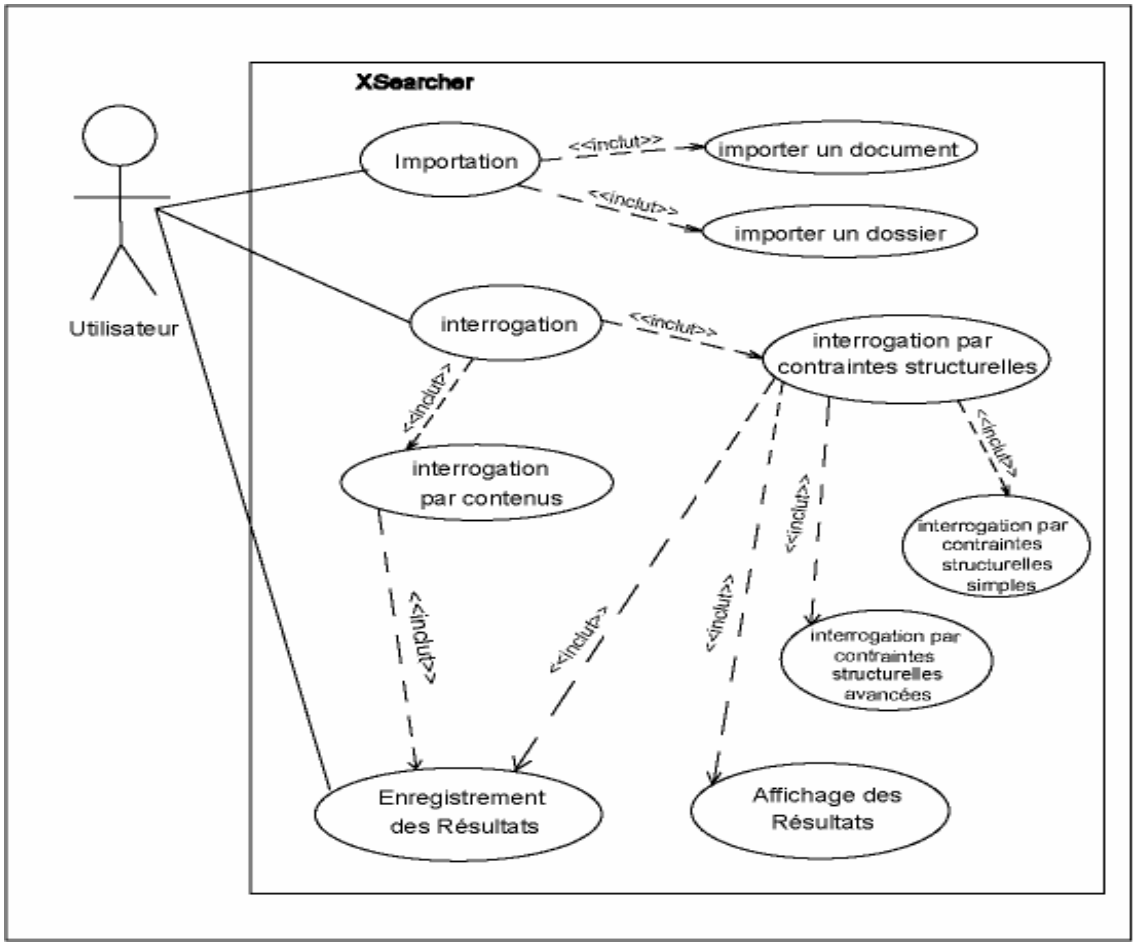

Figure 1 : Architecture du système par le diagramme de cas d'utilisations d'UML

3.2- Implémentation du Système XSearcher

Afin de nous aider dans le développement de notre système, nous avons étudié en détail plusieurs implémentations du langage XQuery (toutes recommandées par le W3C, ce qui garantie qu'elles respectent la spécification XQuery définie par le W3C) telles que Quip de Software AG, Qizx/open de Xavier C. Franc, Saxon de Saxonica, XQuark du XQuark Group en collaboration avec l'université de Versailles Saint-Quentin, XHive/DB de X-Hive, XStream DB de Blue Stream Database Software Corporation, Derby de GAEL, Stylus studio de Sonic Software, etc.

Suite à cette étude, nous avons opté pour l'implémentation XHive/DB et ce, pour les raisons suivantes :

- XHive/DB implémente la spécification de XQuery datant du 12 Novembre 2003.

- X-Hive/DB permet de définir une base de données XML native permettant ainsi un stockage et une manipulation rapides des documents XML.
- X-Hive/DB permet aux programmeurs d'accéder au contenu de la base de données via une API java.
- XHive/DB implémente la plupart des fonctions XQuery définies par le W3C [5] dans le document « XQuery 1.0 and XPath 2.0 functions and operators » [7].

L'interface principale du système XSearcher se présente comme suit :

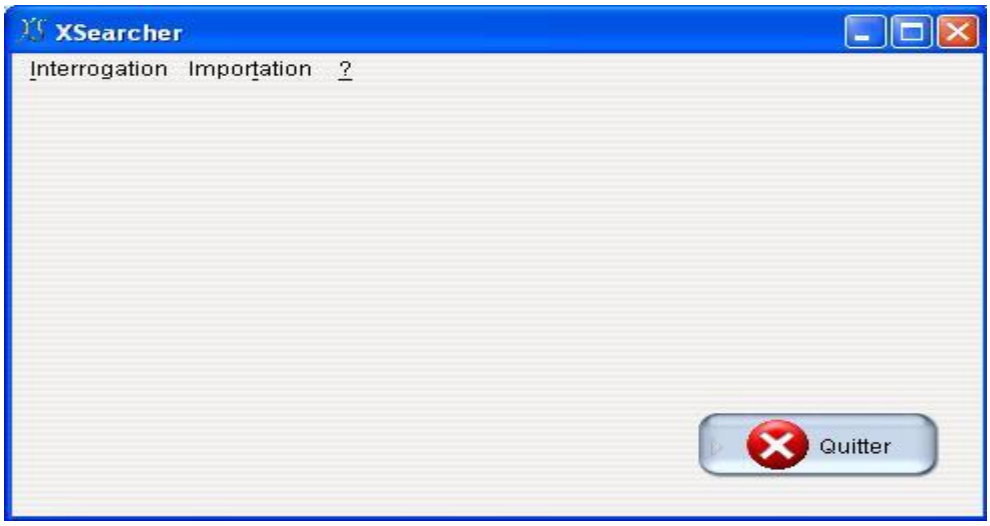

Figure 2 : Interface principale du Système XSearcher

Cette interface se compose de trois menus : « Importation », « Interrogation » et « ? » et d'un bouton « Quitter » permettant de quitter l'application.

a- Le menu Importation

Le menu Importation (ou module Importation) permettra d'importer des documents XML dans la base de documents gérée par le système Xsearcher. Ces documents doivent être valides par rapport à un schéma XML donné avant de les importer de telle sorte que seuls les documents valides soient présents dans la base.

"XSearcher" propose deux types d'importation : nous pouvons soit importer un document XML, soit importer un dossier contenant plusieurs documents XML.
Dans le cas de l'importation d'un document, le mécanisme d'importation consiste à créer une copie du document XML sélectionné par l'utilisateur et à insérer cette copie dans la base de documents. C'est au module d'importation de se connecter à la base de

documents et d'y insérer la copie et ce, de manière totalement transparente à l'utilisateur.
Dans le cas de l'importation d'un dossier, le mécanisme d'importation consiste à itérer la procédure que nous venons de décrire à tous les documents du dossier sélectionné par l'utilisateur. Dans les deux cas, l'utilisateur effectue sa sélection via une interface graphique qui transmettra au module d'importation le chemin d'accès au document XML ou au dossier que l'utilisateur souhaite importer. Nous montrons dans ce qui suit une des interfaces graphiques d'importation de documents et qui est celle de l'importation d'un dossier.

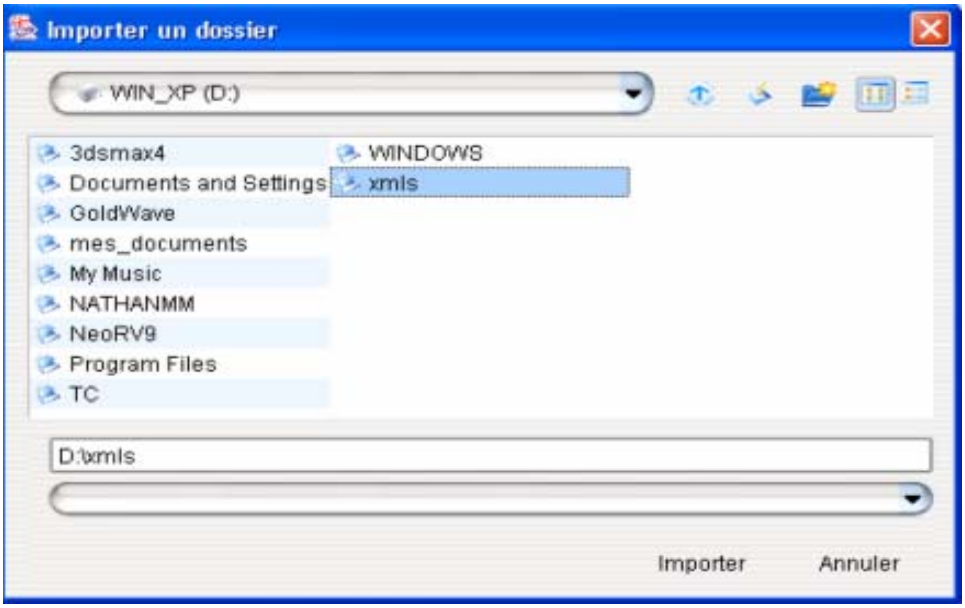

Figure 3 : Importation d'un dossier

b- Le menu Interrogation

Le menu Interrogation (ou module d'interrogation) se charge de générer de manière automatique des requêtes XQuery, à partir de la spécification des besoins des utilisateurs, et de les exécuter.

Le module d'interrogation permettra de faire deux principaux types d'interrogation :

- **Interrogation par contenus**,
- **Interrogation par contraintes structurelles sur le document.**

b.1- Interrogation par contenus : dans ce type d'interrogation, l'utilisateur saisi un ou plusieurs mots que le module d'interrogation utilisera pour rechercher les documents

contenant les mots saisis. La recherche s'applique à tout le document (dans le contenu de tous les éléments du document). Ce module retourne la liste des documents XML contenant ces mots.

La commande « Interrogation par Contenus » du menu « Interrogation » donne accès à l'interface suivante :

Figure 4 : Interrogation par contenus

Cette fenêtre est composée d'une zone de saisie dans laquelle on saisie les mots à rechercher. Juste en dessous ce trouve deux boutons radio permettant de définir la méthode de sélection. Le bouton « Tous les contenus » indique que le(s) document(s) résultant(s) de la recherche est celui qui contient tous les mots saisis. En revanche, le bouton « Au moins un des contenus » indique que le(s) document(s) résultant(s) de la recherche doit contenir au moins un des mots saisis.
Au bas de la fenêtre se trouvent trois boutons. Le bouton « Fermer » permet de fermer cette fenêtre. Le bouton « Aide » fournit de l'aide concernant l'utilisation de cette fenêtre. Enfin, le bouton « Lancer la Recherche », qui n'est actif que lorsque des contenus sont saisis, permet d'interroger les documents de la base de documents avec les contenus saisis. Une fois la recherche lancée et effectuée, l'interface suivante s'affiche :

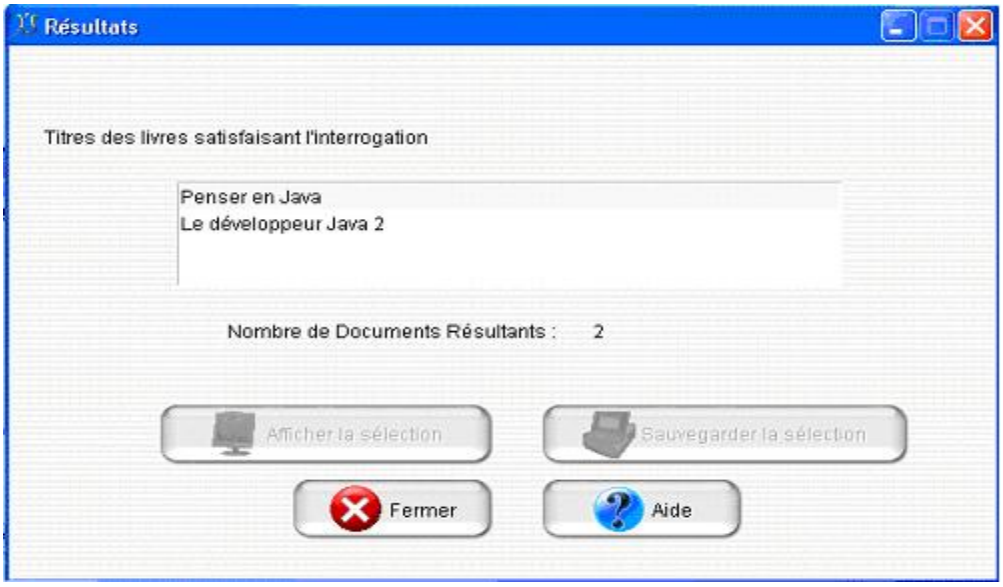

Figure 5 : Affichage des résultats de l'interrogation par contenus

Sur cette fenêtre nous retrouvons une zone de liste dans laquelle sont affichés les titres des livres satisfaisant la requête. Dans le cas où il n'y a aucun résultat, cette liste sera vide.

Le bouton « Afficher la sélection » permet d'afficher le livre dont le titre est sélectionné dans la liste, dans le navigateur Microsoft Internet Explorer. Cet affichage ressemble à ceci :

Figure 6 : Affichage d'un livre dans le navigateur

Enfin, le bouton « Sauvegarder la sélection » permet de sauvegarder le livre dont le titre est sélectionné dans la liste en tant que nouveau document XML. Un clic sur ce bouton lancera la boite de dialogue suivante :

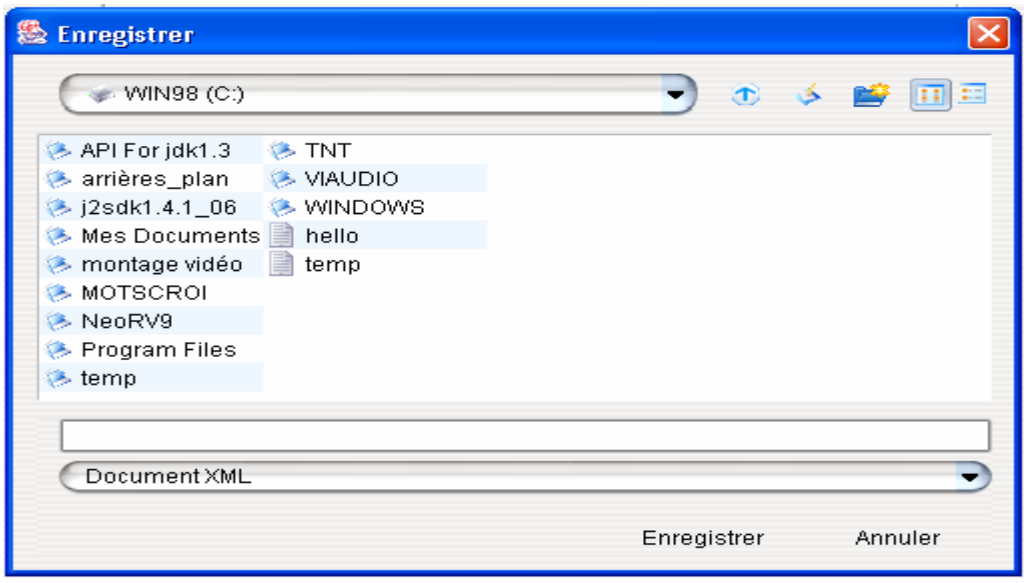

Figure 7 : Sauvegarde des résultats

Dans cette boite de dialogue, nous devons entrer un chemin d'accès pour le document à sauvegarder ou sélectionner un document existant afin de l'écraser.

b.2- Interrogation par contraintes structurelles sur le document : pour ce type d'interrogation, l'utilisateur pourra soit effectuer une recherche par contenus mais en spécifiant l'élément (partie du document) sur lequel il souhaite effectuer sa recherche et c'est ce que nous appellerons le mode «interrogation par contraintes structurelles simples » ; ou bien effectuer une recherche plus fine et plus précise concernant la structure en saisissant d'éventuelles valeurs pour certains éléments du document afin de réduire l'espace de recherche. C'est ce que nous appellerons le mode « interrogation par contraintes structurelles avancées ». Les résultats retournés par ce type d'interrogation sont les éléments des documents contenants les informations recherchées.
Dans le cas de l'interrogation par contraintes structurelles simples, l'interrogation ne se fera que sur l'élément sur lequel l'utilisateur souhaite effectuer sa recherche.

La commande « Interrogation par Contraintes Structurelles Simples » du menu « Interrogation » donne accès à l'interface suivante :

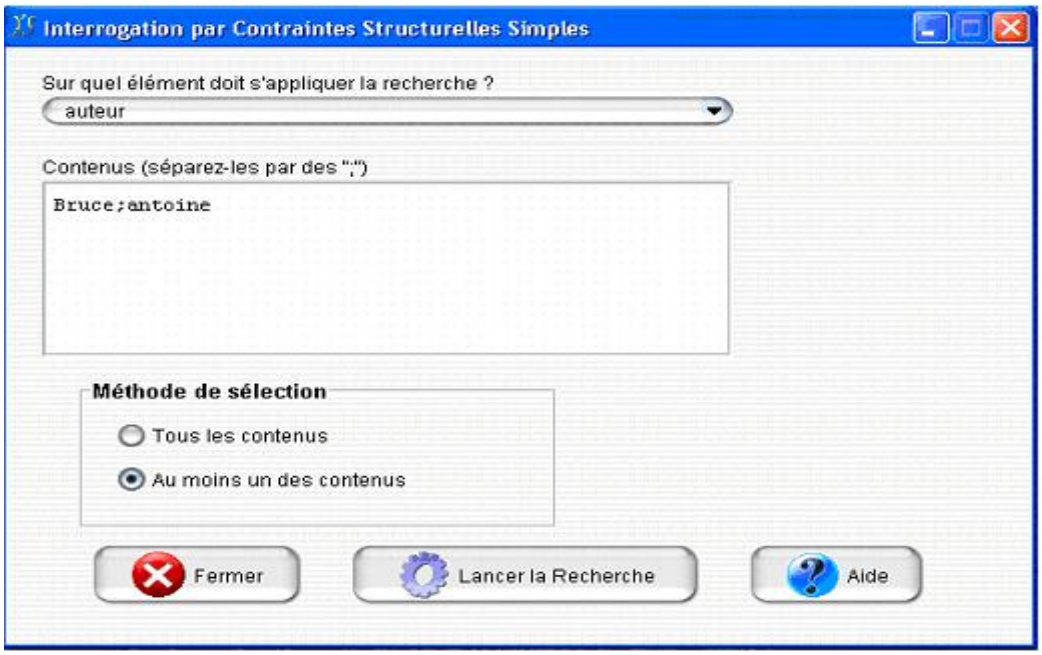

Figure 8 : Interrogation par contraintes structurelles simples

Dans le cas de l'interrogation par contraintes structurelles avancées, par contre, l'utilisateur commencera par indiquer le type de l'information qu'il recherche (par exemple l'année d'édition d'un livre dans une collection de livres ou un numéro de sécurité sociale d'un employé dans une collection d'employés). De là, il disposera de champs lui permettant d'affiner sa recherche afin de réduire l'espace de recherche en saisissant des informations. Par exemple, si l'utilisateur recherche le titre d'un livre, il pourra spécifier (s'il le désire) des informations telles que l'année d'édition du livre, le ou les auteurs du livre, la maison d'édition de ce livre, etc. ce qui permettra d'affiner sa recherche et par conséquent avoir un résultat plus restreint et plus précis. Toujours dans ce cas, l'utilisateur pourra rechercher un élément donné en précisant sa position dans le document XML. Il pourra ainsi rechercher par exemple le titre du 3^{e} paragraphe du 2^{e} chapitre d'un livre donné.

La commande « Interrogation par Contraintes Structurelles Avancées » du menu « Interrogation » donne accès à l'interface suivante :

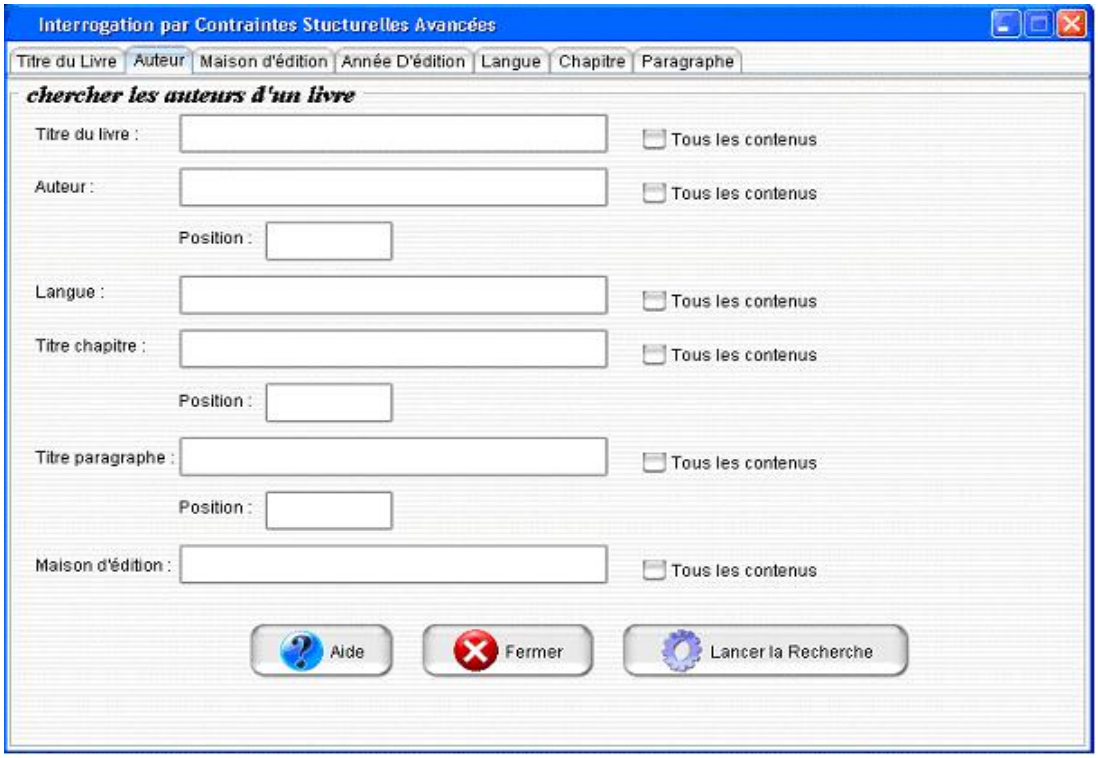

Figure 9 *:* Interrogation par contraintes structurelles avancées

Cette fenêtre se compose de sept onglets représentant les éléments sur lesquels peuvent s'effectuer les recherches (titre du livre, auteur, maison d'édition,…). Ainsi par exemple, sur l'illustration ci-dessus, nous avons sélectionné l'onglet « Auteur » pour rechercher un auteur donné.

En dessous, se trouvent des zones de saisie où nous pouvons saisir des informations relatives aux éléments auxquels elles réfèrent et ce, afin d'obtenir des résultats de recherche plus fins et plus précis.

Les cases à cocher « Tous les contenus », comme leur nom l'indique, feront en sorte que les résultats contiennent tous les mots saisis. Si cette case est décochée, les résultats devront contenir au moins un des mots saisis pour être retenus.

Le bouton « Lancer la Recherche » permet d'interroger les documents de la base de documents avec les informations saisies. Une fois la recherche effectuée, l'interface suivante s'affiche :

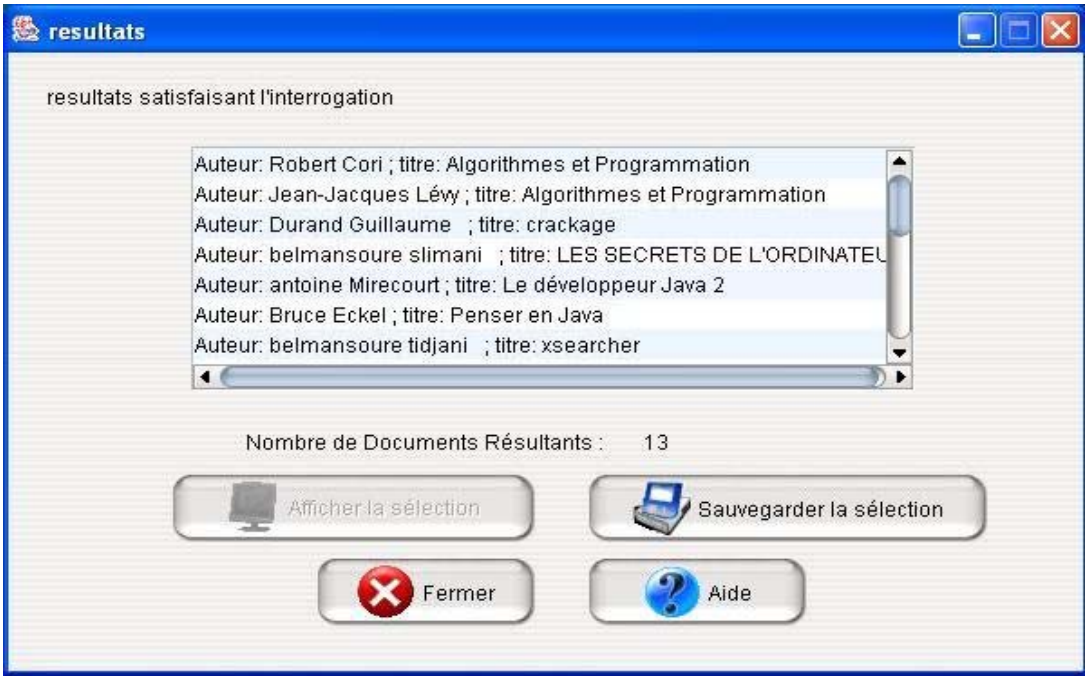

Figure 10 *:* Affichage des résultats pour l'interrogation par contraintes structurelles avancées

Cette fenêtre se compose d'une zone de liste dans laquelle les résultats sont affichés. Les résultats consistent en l'élément sélectionné dans la liste déroulante ainsi que le titre du livre auquel il appartient. Dans le cas où il n'y a aucun résultat, cette liste sera vide.

4- Conclusion

Par ce travail, nous nous sommes penché sur une partie du domaine de la recherche d'information qui est la recherche d'information structurée laquelle porte un intérêt tout aussi grand à la structure qu'au contenu d'un document.

Ce travail nous a amené à concevoir et réaliser le système XSearcher permettant le stockage et l'interrogation de documents au format XML basé sur le langage XQuery.

XSearcher est constitué de trois modes d'interrogation et de deux modes d'importation des documents XML. L'importation consiste à stocker les documents XML dans une base de documents gérée par le système Xsearcher. L'interrogation consiste, quant à elle, à interroger les documents présents dans cette base. L'interrogation peut porter sur le contenu des documents ou sur leur structure. Ainsi, nous pouvons soit rechercher l'occurrence de mots dans un document entier ou dans un élément particulier de sa structure, ou soit effectuer une recherche sur la structure elle-même. XSearcher offre ces possibilités à l'utilisateur à travers une interface graphique et conviviale. De ce fait, la tâche de recherche de l'utilisateur n'est pas complexifiée par la syntaxe d'un langage qui est généralement difficilement accessible pour les novices.

Le Système XSearcher ainsi conçu et réalisé, fournit des résultats fins, précis et pertinents pour des requêtes fines et précises.

Le Système XSearcher a ainsi permis de montrer que la connaissance de la structure des documents est une ressource additionnelle qui améliore la pertinence du système de recherche d'information par un accès plus efficace aux documents susceptibles d'intéresser l'utilisateur. Ceci dit, et afin d'améliorer les performances du système XSearcher, il est nécessaire de généraliser le système afin qu'il permette d'interroger n'importe quel type de documents XML, conformes à n'importe quel schéma XML.

Bibliographie

[1] Initiation à XML – avec 3 études de cas détaillées
D. Hunter, C. Cagle, D. Gibbons, N. Ozu, J. Pinnock, P. Spence
Editions Eyrolles – 2001
ISBN = 2-212-09248-2

[2] XML in a nutshell – manuel de référence – 2e édition en français
Elliotte Rusty Harold & W. Scott Means
Editions O'REILLY – 2002
ISBN = 2-84177-223-3

[3] XQuery: A Query Language for XML
W3C Working Draft 15 February 2001
www.w3.org/TR/2001/WD-xquery-20010215/

[4] XQuery 1.0: An XML Query Language
W3C Working Draft 12 November 2003
www.w3.org/TR/2003/WD-xquery-20031112/

[5] About the World Wide Web Consortium
http://www.w3.org/Consortium/

[6] Extensible Markup Language (XML) 1.0 (Second Edition)
W3C Recommendation 6 October 2000
http://www.w3.org/TR/2000/REC-xml-20001006

[7] XQuery 1.0 and XPath 2.0 Functions and Operators
W3C Working Draft 12 November 2003
http://www.w3.org/TR/2003/WD-xpath-functions-20031112/

[8] XML Query Use Cases
W3C Working Draft 12 November 2003
http://www.w3.org/TR/2003/WD-xquery-use-cases-20031112/

[9] Techniques d'apprentissage pour le traitement d'informations structurées : Application à la recherche d'information.
Benjamin Piwowarski, thèse de doctorat de l'université Paris 6

[10] Toward a Structure Information Retrieval System on the Web
Mathias Géry, Jean-Pierre Chevalet
Equipe MRIM, Laboratoire CLIPS-IMAG, France, 2001

[11] Searching and Browsing Collections of Structural Information
Jens E. Wolff, Holger Florke
Institut d'informatique III, Université Bonn, Germany, 2000

[12] A model for structured document retrieval
Mounia Lalmas, Ian Ruthven
University of Glasgow, 1997

[13] La recherche d'information dans une base de documents structurés en XML
Bessai Fatma-Zohra
Séminaire EdiCulture, XML et l'Echange de Données Informatiques, Ecole Nationale d'Administration, ENA, Hydra, Alger, Avril 2003