Quatrième Partie

4. L’accès à distance aux manuscrits

Le noyau de 173 champs, dont 73 possèdent en moyen 2 ou 3 attributs que nous avons créé pour définir la structure des documents décrivant les manuscrits arabes numérisés, nous a permis de construire une interface permettant l'indexation, la recherche et l'affichage de ces documents. Le logiciel SDX (Système Documentaire en XML) a été choisi, parmi d'autres, pour effectuer cette tâche particulière. Ce chapitre est donc consacré en premier lieu à la description de la structure et de la fonctionnalité du logiciel SDX et, en deuxième lieu, à la description de la base de données et de son interface, et à leur mode d'utilisation.

4.1. Le logiciel SDX (Système Documentaire en XML)

SDX est un outil de traitement et de la diffusion des informations patrimoniales, développé depuis l'année 2000 sur l'initiative de la mission de la recherche et de la technologie du ministère de la culture et de la communication en France. Martin Sévigny de la société AJLSM• a développé la plate-forme expérimentale SDX dont l'objectif est de faciliter la diffusion de bases de documents XML sur l'Internet, ainsi que leur gestion. « **SDX** est un système ouvert (composé entièrement de logiciels libres) et modulaire (de nouveaux services peuvent être ajoutés à l'application initiale) qui peut s'adapter à tout type de documents XML »[118]**.** De nouvelles versions sortant régulièrement, le SDX-2 est une version plus avancée de SDX. Le figure[119] ci-dessous est un aperçu global de SDX-2.

• **AJLSM** est une société spécialisée dans la mise en place de solutions de gestion de l'information électronique, basées sur XML et SGML voire (http://www.ajlsm.com)

[118] **Dalbéra, Jean-Pierre** et **Martin Sévigny.** SDX : Système Documentaire en XML Une plate-forme ouverte et modulaire pour la diffusion et la gestion d'ensembles de documents XML. http://debussy.culture.fr/docs/sdx.html

[119] http://sdx.culture.fr/sdx/journeeSDX/pres/brihaye/brihaye.html#d0e28

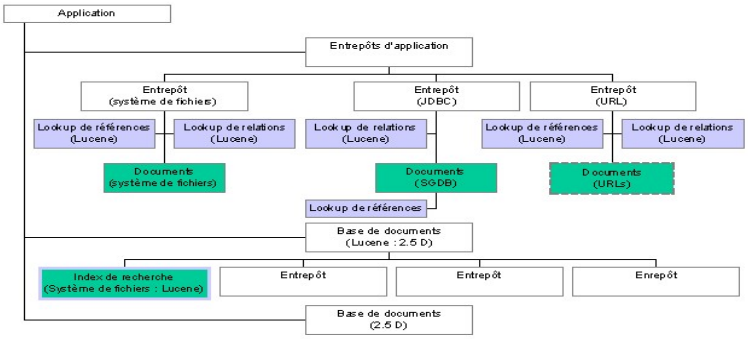

Figure n°. 60 : Aperçu global de SDX-2

4.1.1. La fonctionnalité de SDX

Il faut mentionner que SDX ne possède aucune interface graphique. C'est une application qui regroupe plusieurs autres applications développées en « open-sources ». Elle a son propre langage de programmation (XSP) pour unifier les commandements nécessaires à l'utilisation des différentes applications. Elle se compose de plusieurs logiciels qui permettent de *publier,* d'*indexer* et de *rechercher* des documents XML. Par exemple :

- **Cocoon**[120] : pour la publication, le SDX utilise le logiciel Cocoon. Cocoon est un outil qui aide à l'utilisation de XML et de XSLT pour des applications de serveur.
- **XML:DB** : SDX permet d'indexer des documents XML afin de faciliter leurs recherches. Pour indexer les documents, SDX utilise le base de données XML « XML:DB »
- **eXist** est une base de données qui propose un mécanisme d'indexation et de recherche en plein texte par mots clés.

«La force de SDX est donc de proposer une indexation complètement personnalisable qui, dans un cas extrême, pourrait se passer totalement du contenu du document original. On pourrait ainsi la réduire à des métadonnées par exemple ; c'est une pratique courante

[120] Pour savoir plus sur Cocoon voire l'adresse émail : http://xml.apache.org/cocoon/index.html

chez le bibliothécaire et les archivistes qui n'indexent pas le contenu des documents dont ils ont la charge mais les moyens d'accéder à ces documents » [121]

4.1.2. Le recherche des documents par SDX:

Le logiciel SDX est structuré de manière à gérer que de l'information en XML. En conséquence, il peut indexer n'importe quel type de documents XML et il permet la configuration des champs de recherche de façon très souple bien adaptée aux informations gérées et aux besoins des utilisateurs. Par contre, toutes les bases de données qui n'existent pas en format XML peuvent être transformées rapidement en XML. La souplesse de SDX facilite une recherche simultanée dans plusieurs bases de documents, qu'elles soient définies auprès d'une même application ou dans différentes applications.

Pour la recherche des documents, il est possible sur SDX d'effectuer les deux types de recherche : une recherche textuelle ou par champ. « De plus, SDX offre non pas *une* mais *autant* d'indexations que l'on souhaite tout en se proposant d'offrir à l'utilisateur une syntaxe de recherche simple, à tout le moins plus simple que celle de XPath, SQL• ou autres langages de requête»[122].

4.1.3. L'architecture de SDX-1

Pour la diffusion et la recherche d'une collection de documents XML, le SDX-1 comporte plusieurs composants[123] :

- Un système de stockage (un SGBD relationnel, tel que MySQL[124], InterBase[125], Oracle[126], Hypersonic SQL[127])

---

[121] Brihaye, Pierrick. SDX-2 en tant qu'infrastructure générique pour la conception de systèmes documentaires :. 19pages (P.7) (consulter 03/03/03) http://sdx.culture.fr/sdx/journeeSDX/pres/brihaye/brihaye.html

- Un serveur HTTP (comme celui de Tomcat, Apache ou autre)
- Un moteur de servlets Java (Tomcat ou autre)
- Un processeur XSLT pour transformer l'XML généré en HTML (Saxon)
- Un moteur de recherche (Lucene)
- Un générateur de pages XML (Cocoon-1)

Une application type se présentait en quatre volets :

- présentation des documents,
- navigation hiérarchique,
- recherche simple et avancée,
- Administration de la collection de documents.

Dans la version SDX-2 les ingénieurs du logiciel ont essayé de surmonter les inconvénients de la première version de SDX à plusieurs niveaux :

- Le servlet Cocoon pour améliorer la performance des XML, XSLT et XSP.
- Une autre amélioration est apporter à la fonction d'indexation, de recherche et d'affichage de document XML. Pour assurer cette fonction « il faut qu'un serveur Web, un moteur de servlet et SDX soient en marche et correctement configurés pour être prêts à servir les requêtes. La réalisation d'un tel environnement est appelée un *serveur* SDX, c'est-à-dire une instance du serveur en état de marche »[128]
- Le moteur de recherche Lucene qui a été utilisé dans le SDX-1 est théoriquement remplaçable (à terme) mais pratiquement, peu de composants libres pourraient le remplacer. Un autre processeur XSLT peut être choisi sans risque

---

• SQL (Structured Query Language) voir http://www.w3schools.com/sql/sql_select.asp

[122] Brihaye, Pierrick. SDX-2 en tant qu'infrastructure générique pour la conception de systèmes documentaires : Pistes de recherche et développement. http://sdx.culture.fr/sdx/journeeSDX/pres/brihaye/brihaye.html#d0e28 dernière mise à jour le 06/02/2002

[123] **Sévigny**, Martin et Michel Botin. SDX en Quelques mots. http://sdx.culture.fr/sdx/definition.html

[124] http://www.mysql.com/

[125] http://www.borland.com/interbase/

[126] http://www.oracle.com/

[127] http://hsqldb.sourceforge.net/

[128] http://www.nongnu.org/sdx/docs/html/doc-sdx2/fr/presentation/serveur.html

d'incompatibilité, à la condition qu'il soit en Java (ex:Xalan). Xalan-Java est un processeur XSLT qui aide à transformer des documents XML en HTML, texte ou autres types de documents XML « *Xalan-Java is an XSLT processor for transforming XML documents into HTML, text, or other XML document types* »[129].

- Avec SDX-2, les sources de données peuvent être multiples, conduisant au concept générique d'entrepôt sans imposer une base de données extérieures au serveur.

La base de données et l'entrepôt sont deux plates-formes qui doivent être présentées à l'intérieur de SDX.

4.1.4. La base de données dans SDX

4.1.4.1. La base de données une définition

« La base de documents définit donc un ensemble de documents qui peuvent être recherchés. Cet ensemble partage des caractéristiques communes, en particulier un certain nombre de champs et leurs caractéristiques. On pourrait donc dire qu'une base de documents est définie fondamentalement par une liste de champs et par le contenu de ces différents champs, produit lors de l'indexation de documents »[130].

Un des avantages de SDX est qu'il n'impose pas de restriction sur les documents qu'il indexe au sein d'une base. Par contre il est possible d'indexer des documents XML respectant différents schémas ou DTD au sein d'une même base.

4.1.5. Le SDX et le multilinguisme

SDX accorde une grande importance à la langue et c'est pourquoi il est présenté comme un outil multilingue. « Autrement dit SDX est une plate-forme qui permet de réaliser des applications de recherche multilingues, cela ne signifie pas que SDX prévoit déjà tous les

---

[129] http://xml.apache.org/xalan-j/

[130] **Sévigny**, Martin et **Frédéric** Glorieux. Les bases de documents, http://www.nongnu.org/sdx/docs/html/doc-sdx2/fr/presentation/bases.html (dernier mise à jour, 2002/09/25)

outils nécessaires pour traiter correctement toutes les langues souhaitées par un développeur »[131].

La langue dans SDX est construite autour de deux technologies (Java et XML) qui chacune a sa propre conception. C'est une « combinaison d'un code de langue ISO-639, d'un code de pays ISO-3166 (optionnel) et d'une variante (optionnelle). Pour représenter ces langues dans un contexte XML, l'attribut XML:lang est utilisé pour les deux premières parties et un attribut "variant" est utilisé pour le troisième. Exemple :

```
<sdx:firld xml:lang= "fr-CA"variant="qc"…/>»[1]
```

Mais pour la langue d'interface, le code de pays est identifié à l'aide de l'attribut xml:lang. Pour construire les pages web dynamiques le SDX utilise en général le langage XSP dans laquel on trouve un élément qui s'appelle <sdx:page> qui permet d'utiliser par la suite l'un ou l'autre des services SDX. Exemple[132] :

```
<sdx :document xml :lang="fr-CA">
……
<sdx:results…>
…..
</sdx:results>
</sdx:document>
```

L'exemple ci-dessus montre que le document XML aura, à sa racine, un élément <sdx:document> et un attribut xml:lang qui identifie la langue souhaitée pour l'interface. Cet attribut est toujours présent pour créer des interfaces multilingues.

Pour la langue de l'utilisateur et dans l'interface de gestion des utilisateurs de SDX, il est possible d'associer une langue à l'utilisateur. «Les applications qui ont leur propre interface de gestion des utilisateurs peuvent, de leur côté, associer une langue avec la méthode de leur choix, et SDX pourra tout de même utiliser cette langue pour indiquer la langue d'interface. »[133]

---

[131] **Sévigny**, Martin et **Frédéric** Glorieux. SDX et le multilinguisme. http://www.nongnu.org/sdx/docs/html/doc-sdx2/fr/presentation/multilinguisme.html (dernier mise à jour, 2002/09/25)

[132] **Sévigny**, Martin et **Frédéric** Glorieux. SDX et le multilinguisme. http://www.nongnu.org/sdx/docs/html/doc-sdx2/fr/presentation/multilinguisme.html (dernier mise à jour, 2002/09/25)

[133] **Sévigny**, Martin et **Frédéric** Glorieux. SDX et le multilinguisme http://www.nongnu.org/sdx/docs/html/doc-sdx2/fr/presentation/multilinguisme.html (dernier mise à jour, 2002/09/25)

4.1.6. L'analyseur de mot

L'objectif d'un analyseur de mot est de transformer le contenu textuel dans le champ de type *word*. Après l'analyse, chaque mot individuel est stocké dans l'index pour distinguer les mots minuscules et pour supprimer les diacritiques. Mais il est difficile de trouver un analyseur de mot qui fonctionne bien pour toutes les langues car cela dépend de la langue de texte. Pour cette raison, le SDX permet au développeur d'application d'utiliser des analyseurs de mots différents pour tous les champs des bases de documents. «Ces analyseurs peuvent être soit inclus dans SDX, soit fournis par le développeur lui-même ; un analyseur de mots est tout simplement une classe Java qui étend (directement ou indirectement) la classe fr.gouv.culture.sdx.search.lucene.analysis.Analyzer.»[134].

Jusqu'à présent, SDX fournit des analyseurs pour cinq langues seulement, à savoir : la langue *chinoise*, la langue *allemande*, la langue *anglaise*, la langue *français*e et la langue *russe*. Mais il a la possibilité accepter d'autres contributions pour d'autres langues car en structure interne, il fournit un mécanisme de configuration des analyseurs. Le fichier de configuration pour chaque langue se trouve sous l'adresse: sdx/resources/conf/analysis exemple[135] :

```
< ?xml version= "1.0" encoding="ISO-8859-1" ?>
<french useStopWords="true" keepAccents="false">
<stopWord>
<stopWord>le</stopWord>
<stopWord>la</stopWord>
<stopWord>les</stopWord>
</stopWord>
</french>
```

La langue arabe est un exemple du manque d'outils nécessaires pour la traiter au sien de SDX. Donc la base de donnée créée pour les manuscrits arabes sur SDX ne fournit par une recherche par la langue arabe, espérons avoir le moyen d'intégrer les outils nécessaires pour effectuer cette tâche le plutôt possible.

---

[134] **Sévigny**, Martin et **Frédéric** Glorieux. SDX et le multilinguisme. http://www.nongnu.org/sdx/docs/html/doc-sdx2/fr/presentation/multilinguisme.html (dernier mise à jour, 2002/09/25)

[135] **Sévigny**, Martin et **Frédéric** Glorieux. SDX et le multilinguisme. http://www.nongnu.org/sdx/docs/html/doc-sdx2/fr/presentation/multilinguisme.html (dernier mise à jour, 2002/09/25)

4.2. La base de données actuelle

Notre intention, dans cette partie, est de décrire la base de données hébergée sur le serveur de L'enssib et l'interface d'accès aux manuscrits arabes. Le but principal de cette base est de permettre aux utilisateurs d'effectuer des recherches parmi l'ensemble des champs sélectionnés. Une quarantaine de champs ont ainsi été sélectionnés parmi le 173 définis dans la DTD. « Le site devait donc permettre, pour le résultat d'une recherche, l'affichage structuré de documents descriptifs stockés dans un format sans mise en forme (XML) et l'affichage des images numérisées décrites dans le document XML »[136]. La création de site Web a vu le jour grâce à un travail coopératif avec Guillaume Bourgois un étudiant de l'IUT Informatique de l'université Lyon 1, lors de son stage effectué à l'enssib en vue d'être utilisé plus tard à une plus grande échelle.
Le page d'accueil si dessous montre les deux façons de recherche offertes par la base: Une recherche simple et une recherche avancée.

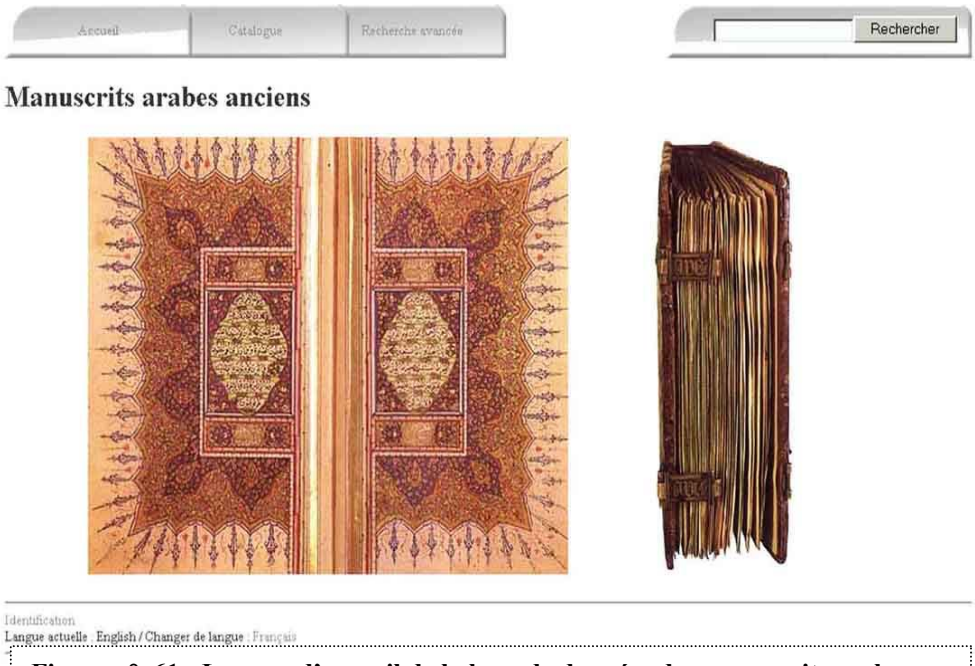

Figure n°. 61 : Le page d'accueil de la base de données des manuscrits arabes

[136] Bourgois, Guillaume. Création d'une interface web pour l'indexation, la recherche et la consultation de manuscrits arabes anciens numérisés- Rapport de stage effectué à l'enssib du avril au juin 2003. 37pages. (p.9)

La recherche simple : elle ouvre une fenêtre cohérente à une séries des sept petites icônes. Par ces icônes il est ainsi possible de consulter:

- La liste des titres des manuscrits disponibles,
- La liste des auteurs dont un ou plusieurs manuscrits ont été indexés dans la base,
- La liste des copistes qui ont recopié, à partir d'un manuscrit original, un ou plusieurs des manuscrits indexés dans la base de données,
- La liste des sujets traités dans les manuscrits,
- La liste des styles d'écriture que l'on peut trouver dans les manuscrits,
- La liste des pays où sont stockés les manuscrits papier,
- La liste des institutions qui conservent les manuscrits papiers.

En cliquant sur une de ces petites icônes, une liste fait apparaître sur l'écran les informations correspondes (exemple : la liste des auteurs). Chaque élément apparaissant dans une liste contient un lien vers l'affichage d'un des documents si l'élément ne fait référence qu'à un seul document, comme par exemple un titre ou vers l'affichage de plusieurs documents dans le cas ou l'élément désigne plusieurs manuscrits, comme dans le cas des listes des styles d'écriture. La figure ci dessous montre l'écran avec les sept petites icônes en haut et un peu plus bas le liste des titres des manuscrits trouvés dans la base. Il est nécessaire de signaler ici que chaque fois que la base sera alimentée par des nouvelles informations, l'affichage de la liste correspondante sera mise à jour.

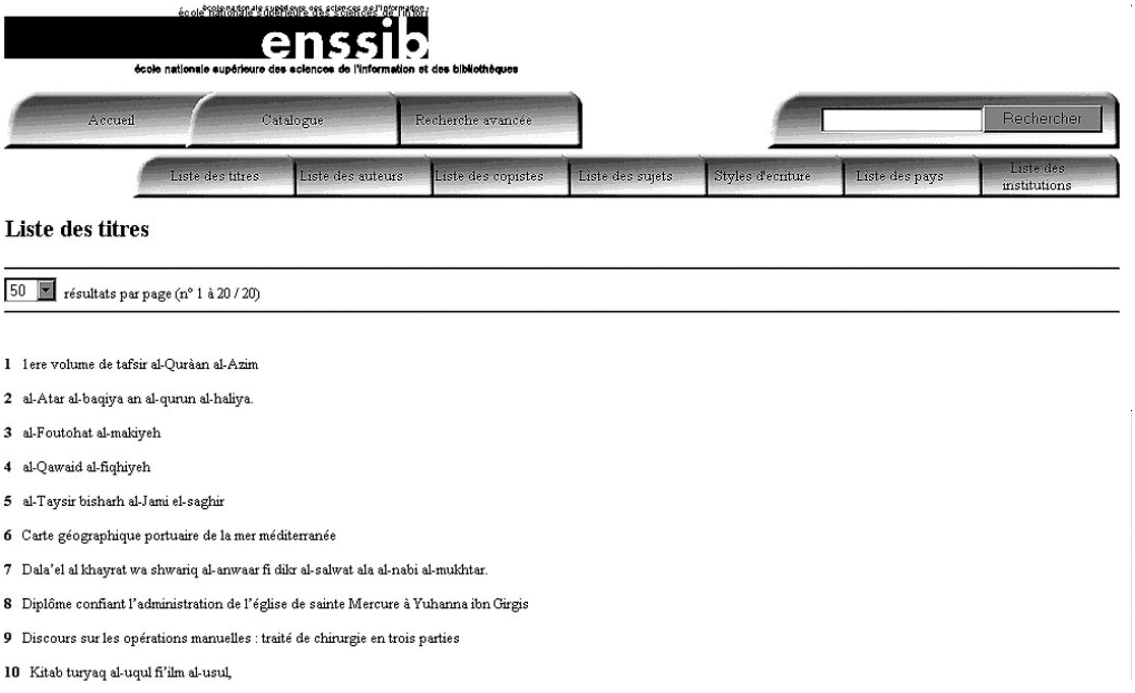

Figure n°.62 : L'interface de la recherche simple dans la base de données des manuscrits arabes

La recherche avancée : la recherche avancée se compose de plusieurs champs qui aide les chercheurs à effectuer une requête composée de plusieurs éléments.

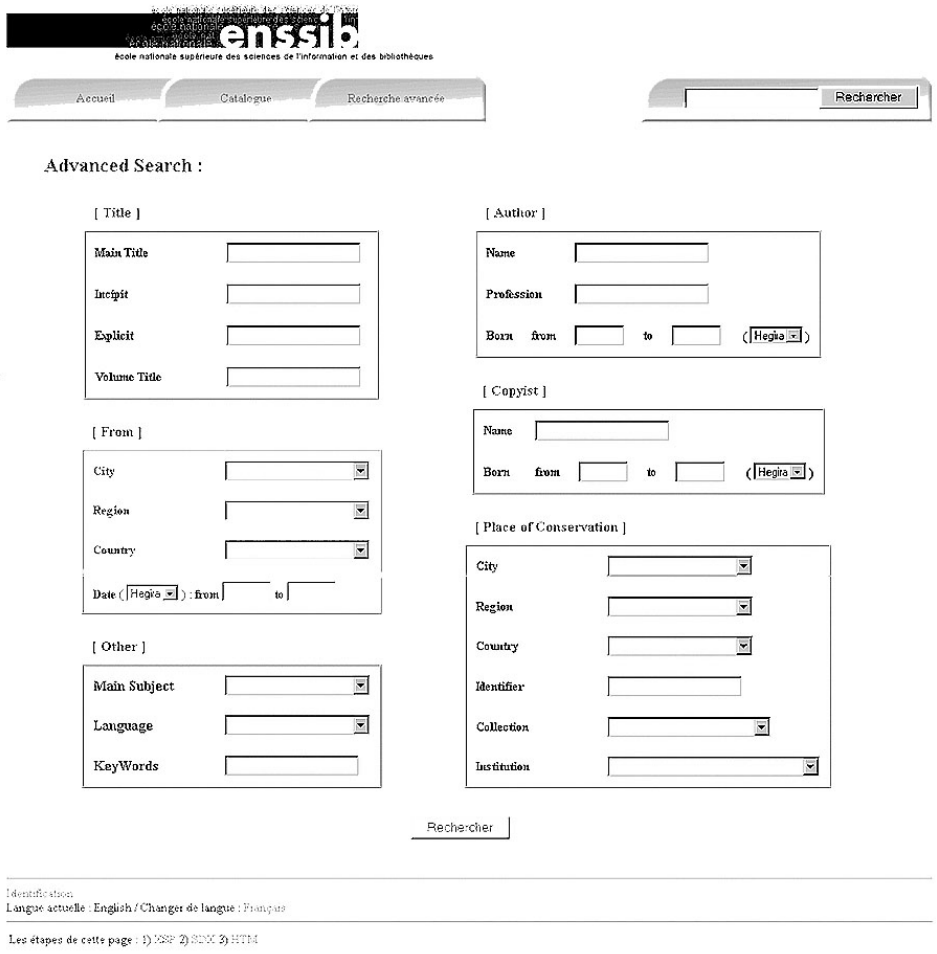

Figure n°.63 : L'interface de la recherche avancée dans la base de données des manuscrits arabe

La figure ci-dessus montre les six grands éléments de recherche : par titre, auteur, copiste provenance, lieu de conservation, et autre. Dans « ***Titre*** » il y a la possibilité d'effectuer une recherche par le titre principale, l'incipit, l'explicite et le titre de volume. Dans «***Auteur*** », il est possible de chercher par le nom d'auteur, sa profession et sa date de naissance et/ou la date de sa mort soit par rapport à l'hégire, soit par rapport à la naissance de JC. Dans «***Copiste*** » on peut chercher par son nom et par la date. La recherche par ville, région, pays et date se fait par l'aide d'éléments trouvés dans «***Provenance***». Il est possible aussi de chercher par le lieu de conservation, (ville,

région, pays, identifiant, collection, établissement). Dans le dernier grand élément « **Autre** » il y a la possibilité de chercher par le sujet, langue et mots clés.
Il est possible aussi d'utiliser la liste déroutante pour choisir le sujet ou la date de publication par exemple. Bien évidemment, une recherche libre est aussi possible sur ces champs (sujet, date, pays, langue etc.).

4.2.1. L'affichage de document

Une fois qu'on a des résultats de recherche, l'affichage de ceux-ci sera donc possible. Sur le format d'affichage, on trouve seulement les champs décrivant un manuscrit donné. Par contre, pour certain champs on trouve un lien avec une URL (adresse de la page dans le base d'image). Ceci arrive où moment où l'image dans la base contient des caractéristiques qui correspondent à ce champ (exemple pour le champ illustration on va avoir un liste des URL pour toutes les images qui contiennent des illustrations dans le manuscrit). En cliquant sur cette adresse, une image sera donc affichée. Il faut signaler ici que tout ce travail de lien a été fait manuellement. Pour automatiser l'indexation, un travail avec l'INSA de Lyon est en cours de réalisation, il sera développé dans le chapitre suivant.

4.2.2. L'affichage de base en multilingue

Jusque présent, le base de données assure un affichage multilingue anglais/français pour un maximum d'éléments. Un affichage en langue arabe serait idéal, mais à cause des problèmes liés à l'application de cette langue dans SDX, ce n'est pas encore réalisable. Un tel travail peut être possible dans le futur proche. Le site étant destiné à un accès mondial (les spécialistes de tout pays doivent pouvoir le consulter), la gestion d'un affichage multilingue est évidement nécessaire.

4.3. Conclusion

SDX est un logiciel utile pour créer des documents XML sur l'internet. Sa flexibilité lui permet d'être très pratique pour le programmeur de créer leur propre base de donnée sans être strictement lié avec un logiciel pré défini qui se trouve sur le marché. Autre avantage de SDX est le fait qu'il soit gratuit. Cet avantage donne beaucoup d'espoir

pour les responsables des bibliothèques ou des centres de recherche de mieux gérer leur bibliothèque avec des dépenses très modestes. Notre base de données pour les manuscrits arabes est consultable par l 'Internet grâce à ce logiciel