heig-vd
Haute Ecole d'Ingénierie et de Gestion
du Canton de Vaud

Tél. : + 41 21 316 98 11
Fax : + 41 21 626 04 67

comem@heig-vd.ch
http://www.heig-vd.ch

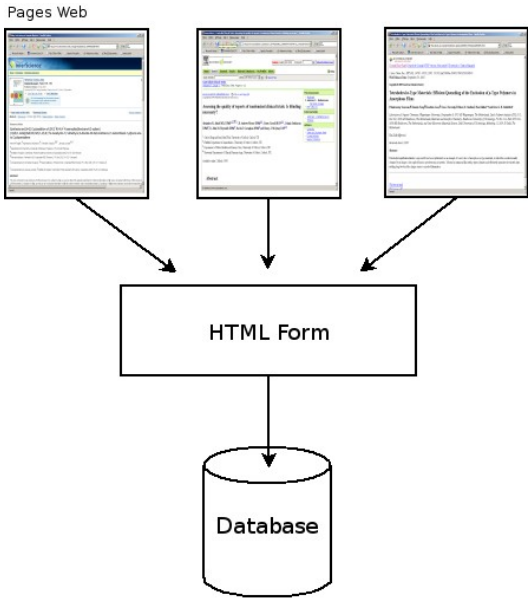
Travail de diplôme 2005

Département Comem+

HEIG-VD
Rue de Genève 55, CP 59
CH - 1000 Lausanne 20

Système d'extraction et d'indexation automatique de pages WEB

Problématique



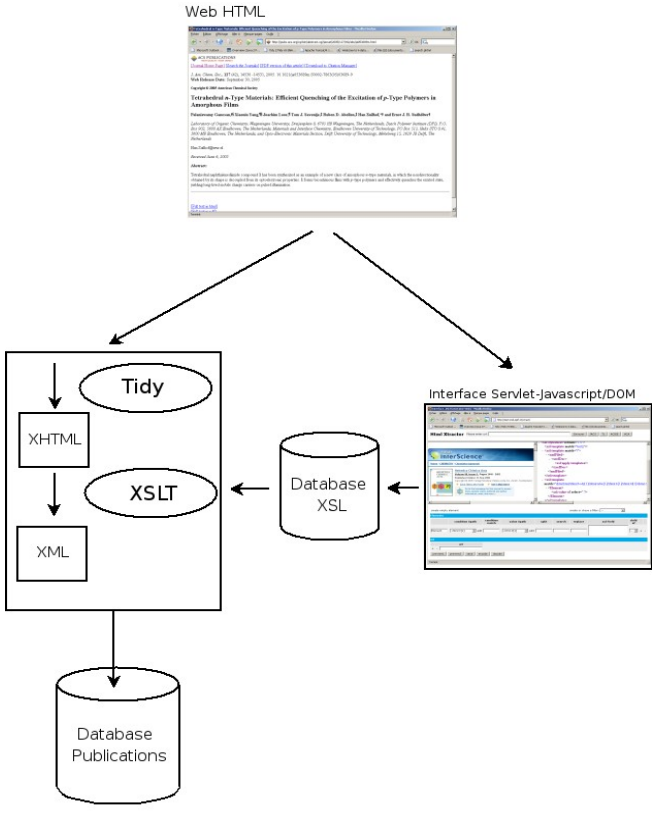
Le Service d'Information Chimique de l'EPFL utilise un outil nommé DBCreator, afin d'indexer dans une base de donnée locale, les publications HTML se trouvant sur Internet. Actuellement, l'indexation nécessite de remplir un à un les champs d'un formulaire HTML. Cette opération demande beaucoup de temps compte tenu du nombre de publications présentes sur Internet.

Objectif

Ce travail de diplôme consiste à développer un outil permettant d'extraire et indexer automatiquement des pages Web. Ainsi, l'indexation est grandement facilitée. L'outil doit être suffisamment général afin de s'intégrer dans d'autres environnements.

Technique

Le principe est de transformer les pages Web en XHTML à l'aide de l'outil Tidy, puis en XML via un filtre XSLT, ce format permettant ensuite une indexation automatique. L'outil développé génère, d'une manière semi-automatique, des filtres XSLT correspondant à des pages Web. Cette génération se réalise dans un environnement de servlets (notamment du framework Cocoon pour les transformations XSLT) et l'interface utilise le couple Javascript/DOM.



Auteur: **Loïc DELACOUR**
Répondant externe: **Luc Patiny**
Prof. responsable: **Nicolas Chabloz**
Sujet proposé par: **Luc Patiny**

Hes·so
Haute Ecole Spécialisée
de Suisse occidentale