Projet d'exploration et extraction de connaissances sur la pollution de l'air depuis une collection de documents publiques

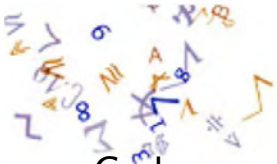
Mihaela Juganaru-Mathieu et Silvia González Brambila

Ecole Nationale Supérieure des Mines SAINT-ETIENNE

UNIVERSIDAD AUTONOMA METROPOLITANA Casa abierta al tiempo Azcapotzalco

le 13 mai 2011

Plan



Justification du projet

Cadre : avoir et traiter des données sur la pollution de l'air de la ville de Mexico

A priori (subjectifs ou pas) : ville polluée, mais nette amélioration, plus de pluies acides

Idéal : Disposer des <u>données fiables</u> et savoir les exploiter/comprendre.

Contexte : la Secretaria del Medio Ambiente de la ville de Mexico (Districto Federal) met à disposition des donnés de mesure au jour le jour (mais 1 valeur/paramètre) et aussi des rapports synthétiques destinés au grand public ou à des spécialistes. Pas d'autres moyens.

But : traiter les documents numériques à disposition et en extraire informations et connaissances.

Collection à traiter

Depuis le site de la Dirección de Monitoreo Atmosférico
`http://www.sma.df.gob.mx/simat2/informaciontecnica/index.php`
une collection des rapports annuels et classifiée en trois catégories :

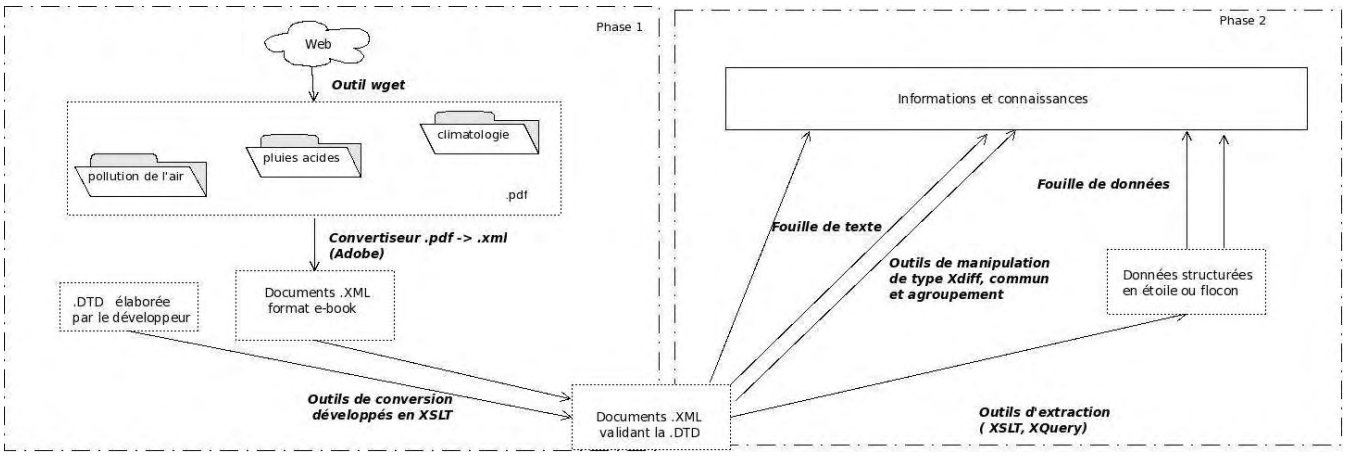
- qualité de l'air (1994–2009)
- pluie acides (1994–1999)
- climatologique (2001-2006)

Documents en format .pdf imprimable (mise en page, texte, données, images, ...).
Chaque document avec 20-40 pages (max 180) et 1-4 Mbyte.

Structuration du projet

Deux grades phases :

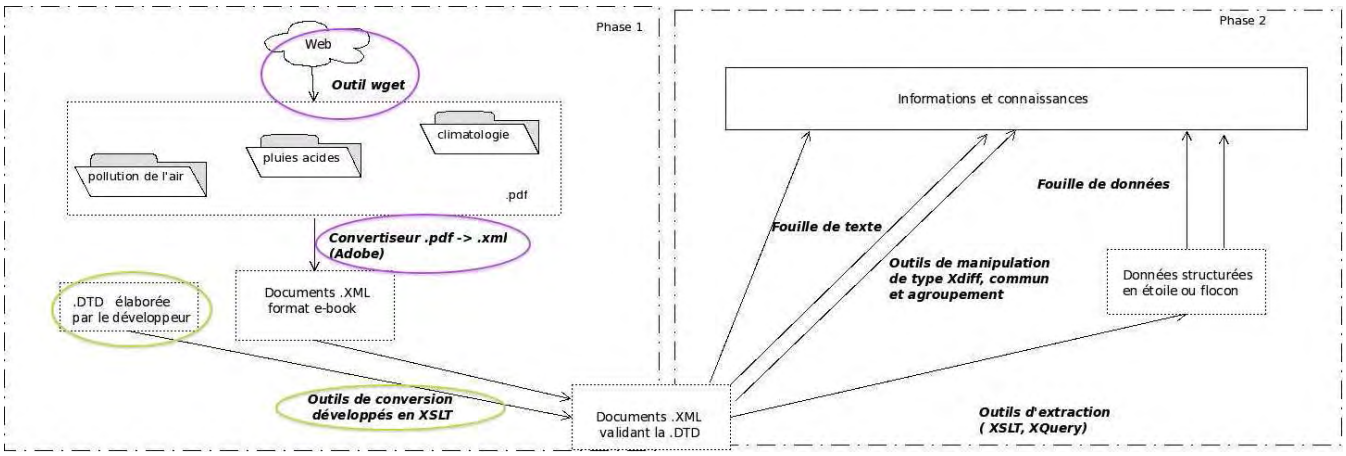
- récupération et mise en forme :
  - obtention, conversion .pdf → .xml (e-book)
  - re-conversion pour filtrer texte et structure, données et signification (XPath, XSLT, XQuery)
  - but : corpus .XML pour la partie texte et pour remplir ultérieurement le datawarehouse (modèle flocon de neige ?)
- fouilles - de texte et de données - pour l'extraction des informations et connaissances



Fouille de texte

La fouille de texte comportera :

- représentation des documents selon le modèle vectoriel (adapté au XML)
- indexation et indexation inverse
- recherche d'information
- mesure de forme et densité
- détection des parties communes (diff) des documents
- recherche de motifs



Perspectives

- le travail est au stade de projet sur les rails
- durée du projet (développement informatique) : 12mois/homme
- aujourd'hui on est à 10%

- besoin futur d'aide de la part des spécialistes en environnement

- traitement de la collection d'images

Merci.

Questions ?