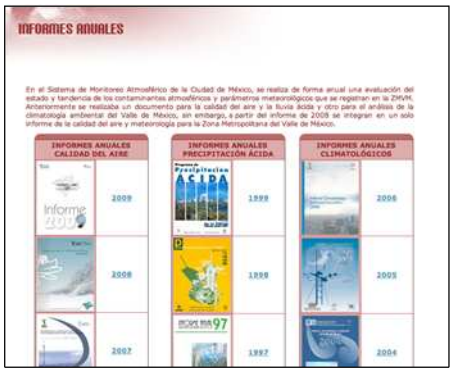
Projet d'exploration et extraction de connaissances sur la pollution de l'air depuis une collection de documents publiques

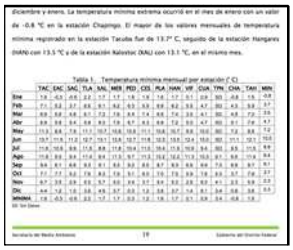
Mihaela Juganaru-Mathieu, Silvia González Brambila***
**Institut Henri Fayol, Ecole Nationale Supérieure des Mines de Saint Etienne,* France mathieu@emse.fr
***Departamento de Sistemas, Universidad Autónoma Metropolitana, Azcapotzalco - Mexico D.F.,* Mexique sgb@correo.azc.uam.mx

Contexte :
trouver et comprendre les documents fiables sur la qualité de l'air sur la ville de Mexico

Source de données :
site web du Secretaría del Medio Ambiente del Gobierno del Distrito Federal (SMA-GDF)
`http://www.sma.df.gob.mx/simat2/informaciontecnica/index.php?opcion=5&opciondifusion_bd=3`

Structure de la colection :
- rapports annuels en format .pdf imprimable
- taille chaque rapport : 20-40 pages (ou plus), 1-4MBytes
- 3 thèmes : qualité de l'air, pluies acides, climatologie

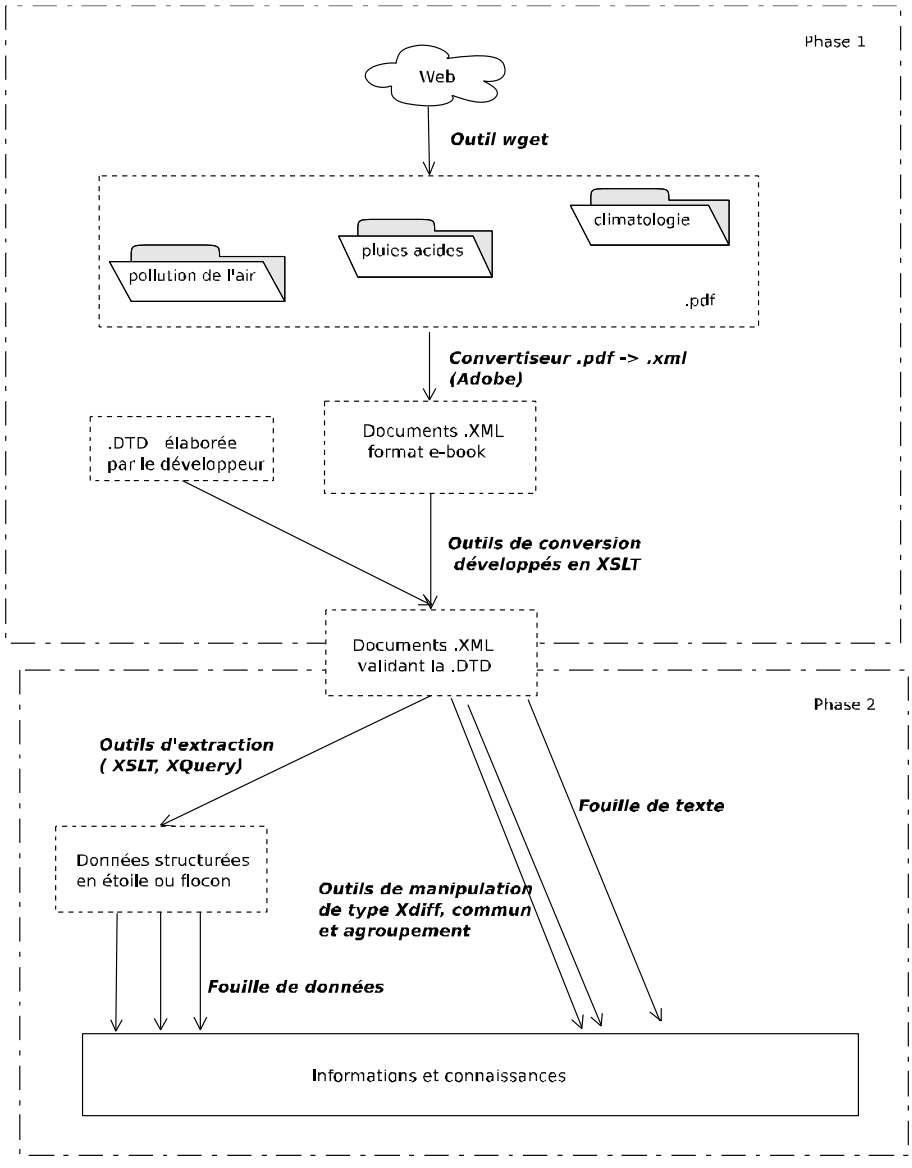




Objectif :
extraction « automatique » des informations et des connaissances à partir de cette collection grand public et fiable (faute de lecture intégrale et complète)

Etapes du projet :
- Téléchargement de la collection
- Nettoyage et structuration des données
- Fouilles de texte (structuré)
- Outils de traitement de documents numériques
- Fouille de données

- Interprétation des résultats



Format de représentation :
Initial : .pdf avec texte (significatif ou pas), données numériques, images
Final : - .xml pour le texte
- données numériques extraites semi automatiquement avec une insertion dans une base de données
- images ignorées

Fouille et traitement de texte :
- Représentation des documents selon le modèle vectoriel (indexation)
- Mesures de densité et de forme
- Recherche d'information
- Détection des parties communes (diff) des documents
- Classification (agroupement) non-supervisé

Travail en cours :
1 mois / homme = environ 10%