haute école neuchâtel berne jura — arc gestion neuchâtel delémont

Hes·so Haute Ecole Spécialisée de Suisse occidentale

Extraction et structuration des données du marché de l'emploi – Gary Criblez

Objectifs

Extraire les données des différents sites internet du marché de l'emploi Romand et leur donner une structure commune. Les informations extraites devront être visualisables

Processus

Interagir

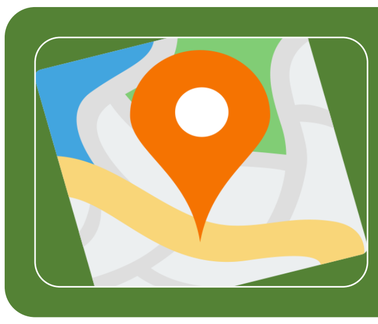
- Accéder aux pages Web
- Simuler l'action d'un utilisateur à l'aide d'un « Web Robot »

Extraire

- Parser les pages Web pour en faire des documents XHTML
- Repérer les balises contenants les données à extraires
- Utilisation des langages XPath, XQuery et les expressions régulières pour la récupération et le nettoyage des données

Structurer

- Regrouper et contrôler les informations extraites
- Stocker dans une base de données



Visualiser

- Visualiser sur une carte les offres d'emploi récupérées.

Web-Harvest

Web-Harvest est un Framework développé en Java. Il permet de faire du «Web Scraping» qui est un processus d'extraction basé sur l'utilisation d'un «Web Robot ».

Celui-ci va s'occuper d'accéder aux pages Web et de transformer, à l'aide d'un parseur, leur contenu HTML en format XTHML afin de pouvoir y appliquer des requêtes XQuery et XPath ou des expressions régulières afin d'en extraire les données et de les intégrer dans une nouvelle structure.