|  |  |
| --- | --- |
|  | **2011** |
|  | IUT de Rodez  Julien Cellier  Xavier Mourgues  Leonhard Hermle |

|  |
| --- |
| **[CAHIER DES CHARGES]** |
| Cahier des charges du projet tuteuré XML exprimant les besoins, les objectifs et les moyens de mise en œuvre du projet. |

Sommaire

[1. Présentation du sujet 3](#_Toc283907053)

[2. Les besoins 3](#_Toc283907054)

[3. Les contraintes 3](#_Toc283907055)

[4. L’univers de travail 3](#_Toc283907056)

[5. Répartition des tâches 4](#_Toc283907057)

# Présentation du sujet

L’objectif du projet est de restituer à partir de sites web existants des données de film (critiques, commentaires, notes, bande annonce, informations diverses…).

Le projet constitue en 3 étapes :

* La récupération de données à travers divers sites web évolutif et dynamique
* Le stockage de ces données dans une base de données Oracle v10.2 (grâce à JAVA)
* La restitution visuel des ces données pour des utilisateurs (en PHP)

Les sites web que nous avons sélectionnés sont : Allocine, Imdb dans un premier temps.

Nous aurons une première phase de découverte visant à étudier les différentes techniques de mise en œuvre d’un tel projet.

Plus tard, dans un souci d’évolution, plus d’information pourront être récupérer et d’autre site dit « source d’information » pourront être rajoutés

# Les besoins

Nous avons besoin de récupérer des données de sites web qui ne sont pas structuré en XML mais en HTML ce qui nous posera problème lorsque voudront traiter ces données. En effet, un site étant évolutif il est possible que sa structure HTML change. Pour parer cet effet, nous devrons réussir à soutirer uniquement les informations nécessaires et non les informations inutiles (publicités …). Pour cela l’utilisation de la technologie DOM sera indispensable.

Dans un deuxième temps, nous aurions besoin d’une machine avec un accès à une base de données Oracle ainsi que qu’un accès libre a Internet (sans proxy bloquant les sites tel qu’Allocine).

# Les contraintes

L’une des premières contraintes que nous avons est l’utilisation d’Oracle v10.2.

D’autre part l’extraction des données se fera par des primitives XSLT mais la structure de données sera en HTML ce qui nous forcera à analyser syntaxiquement ces informations afin d’avoir ce qui se rapproche le plus d’une structure XML.

# L’univers de travail

Nous avons décidé que pour la récupération des données : nous utiliserons un univers java afin de profiter de sa souplesse d’utilisation de ressources tiers tel que le SGBD Oracle via JDBC, le logiciel JTIDY et les primitives XSLT.

Le stockage des données sera partiellement inséré sous format XML et stocké dans une base Oracle.

Pour finir la restitution des données se fera via une PUI en PHP/HTML.

# Répartition des tâches

Nous avons décidé de séparer le projet en 3 parties au sein du groupe.

* M. Leonhard Hermle prendra à sa charge la PUI en PHP/HTML
* M. Cellier Julien s’occupera du stockage en base de données et XML avec interfaçage pour le PHP et le Java.
* M. Mourgues Xavier traitera la récupération des données en Java.