CY Cergy Paris Université
M1 IISC

Projet IED

Le but de ce projet est de concevoir et réaliser un médiateur simple, exploitant différentes sources de données dans le domaine du cinéma.

Le médiateur intègre 3 sources de données :

- Une base de données relationnelle locale, qui sera remplie avec des données extraites du web, en utilisant l'ETL Apache Hop et d'autres outils.
- La source LOD *DBpedia* (https://dbpedia.org), dans laquelle on consultera des informations sur le cinéma : films, acteurs, réalisateurs, producteurs, etc.
- La source *Open Movie Database* (http://www.omdbapi.com/), accessible par des services web REST, qui offre des informations sur des films.

La base de données locale contiendra les informations suivantes sur chaque film : le titre, la date de sortie, le genre, le distributeur, le budget, les revenus aux Etats Unis et les revenus mondiaux. Ces données seront transformées avec Hop pour alimenter la base, à partir des sources suivantes :

- Le fichier *movieBudgets.csv*, fourni sur la page du cours. Il contient les informations sur les films nécessaires pour la base de données, sauf le genre et le distributeur. A part la première ligne qui décrit le type d'information pour chaque film, chaque ligne du fichier contient des informations sur un film, séparées par un caractère tabulation. Attention aux éventuels espaces à la fin des chaines de caractères, qui peuvent fausser les valeurs récupérées.
- Les pages HTML du site www.the-numbers.com, ayant l'URL de la forme http://www.the-numbers.com/market/<année>/genre/<Genre>, où <Genre> peut être Adventure, Comedy, Drama, Action, Thriller-or-Suspense ou Romantic-Comedy et l'année entre 2000 et 2015. Ces pages contiennent dans des tables HTML des informations sur les films du genre en question, y compris le distributeur du film. Elles serviront à compléter l'information sur les films avec le genre et le distributeur.
  Pour chaque genre, un fichier contenant les informations nécessaires pour le médiateur sera extrait de la page HTML, en utilisant un programme Java avec l'API JSoup (https://jsoup.org/). En d'autres termes, pour chaque genre, le programme récupérera toutes les pages HTML pour les années de 2000 à 2015 et produira un seul fichier en sortie par genre. Le format des fichiers extraits est au choix : CSV, XML, etc.

La base de données locale sera remplie à l'aide d'Apache Hop, en prenant tous les films de *movieBudgets.csv* et en complétant le genre et le distributeur pour ceux qui apparaissent aussi dans les fichiers extraits avec JSoup (les autres films resteront sans genre et distributeur renseignés).

DBpedia est accessible en ligne à travers le point d'accès SPARQL http://dbpedia.org/sparql/. Elle contient des informations complémentaires sur les films, notamment le réalisateur, les acteurs et les producteurs du film, nécessaires pour le médiateur. Les requêtes SPARQL adressées à DBpedia peuvent être testées à l'aide de l'interface HTML utilisée en TP, http://dbpedia.org/snorql/.

La source Open Movie Database est consultable à travers un service web de type REST, avec la méthode HTTP GET. Le site http://www.omdbapi.com/ décrit le mode de construction de l'URL d'appel et offre une interface HTML de test. La base contient plusieurs informations sur les films, notamment le résumé, nécessaire pour le médiateur.

Le médiateur utilise un modèle simple, avec une seule relation ***Film***, qui contient tous les attributs mentionnés ci-dessus : titre, date de sortie, genre, distributeur, budget, revenus aux Etats Unis, revenus mondiaux, réalisateur, acteur, producteur, résumé. Comme précisé ci-dessus, le résumé vient de la *Open Movie Database*, le réalisateur, les acteurs et les producteurs viennent de *DBpedia* et le reste de la base de données locale. Le lien entre les informations sur un film venant des trois sources se fait à travers le titre du film.

Le médiateur devra répondre à deux types de requêtes :

1. Etant donné un titre de film, afficher toutes les informations disponibles sur le film : date de sortie, genre, distributeur, budget, revenus aux Etats Unis, revenus mondiaux, réalisateur, résumé, ainsi que la liste des acteurs.
2. Etant donné un nom d'acteur, afficher la liste des films où il a joué. Pour chaque film on affiche le titre, la date de sortie, le genre, le distributeur, le réalisateur et le producteur.

Le médiateur sera programmé en Java, avec une interface textuelle simple, qui permet de choisir le type de requête (par titre ou par acteur), ensuite d'introduire le titre ou l'acteur et d'afficher les résultats. L'accès aux sources de données se fera de la façon suivante :

- *Pour la base de données locale* : en utilisant l'API JDBC.
- *Pour DBPedia* : en utilisant l'API Jena (http://jena.apache.org) pour interroger le point d'accès SPARQL (http://dbpedia.org/sparql/). Plus précisément, on utilisera l'API application de Jena, contenue dans le package `com.hp.hpl.jena.query` (voir http://jena.apache.org/documentation/query/app_api.html).
- *Pour Open Movie Database* : on appelle le service REST par un appel HTTP GET avec l'URI correspondant à la requête, en demandant le résultat en format XML. Pour pouvoir utiliser le service, vous devez souscrire sur le site pour obtenir une clé (API key), qui doit être rajoutée aux paramètres d'appel du service REST.
  Le résultat XML sera transformé en arbre DOM XML en mémoire, ensuite interrogé en XPath pour extraire la partie recherchée (le résumé). Le programme *XPathExemple.java* fourni sur le site du cours montre la façon la plus simple de réaliser ces opérations.

A rendre :

- Un court rapport décrivant le processus d'extraction, de transformation et de chargement des données, ainsi que le fonctionnement du médiateur.
- Le programme Java du médiateur, capable de répondre aux deux types de requêtes.

Une attention particulière sera portée au traitement des informations incomplètes ou ambiguës dans les sources : caractéristiques manquantes, films avec un même titre, etc.

Une démonstration du programme sera présentée par chaque groupe.

Adaptation pour les étudiants travaillant seuls sur le projet

- L'entrepôt local n'intégrera pas les données extraites du site www.the-numbers.com, seul le fichier *movieBudgets.csv* sera utilisé. La base de données locale ne contiendra donc pas le genre et le distributeur du film.