Data Integration

Licenciatura em Engenharia Informática: 2º ano - 2º semestre 2019/2020

Practical Project

Data Integration Using XML

The description of the project is purposely vague, generic and incomplete in some points. What is intended is for students to evaluate the various existing options and choose the one they consider most appropriate for each of the situations they face. All choices must be mentioned and duly justified in the report to be delivered.

1. Objectives

This work intends to create a Java program composed of several Wrappers that obtain data from web and create a XML model to perform searches..
To perform this work, you must use the Java Language, Regular Expressions and the JDOM2 and SAXON APIs studied in practical classes.

2. Description

The objective of the work is to create an integrating application that presents a unified view of information related to films released in cinemas. The data must be obtained from Wikipedia to an XML File:
The link in english is: https://en.wikipedia.org/wiki/Wikipedia

The search page is: https://en.wikipedia.org/wiki/

NOTE: In the ***HttpRequest*** function available in Moodle, this links is the function's first argument. The title of the film to be searched for, is the 2nd argument of the function.

The goal is to wrap from this site information of a given film and organize the data in a XML file with the appropriate structure.

The information to store in the XML file is the following (students can add additional information if relevant):

- title
- image (link)
- year
- Release date in the USA
- country
- director
- cast (main actors)
- duration
- distributed by
- language

- music author
- box office
- …

The XML file must be validated using DTD and XSD.
After that the user can perform searches in the XML file.

Attached is a file with a list of films that can be used to test the application. After finishing the application, students should test with other titles and evaluate the functionality of the implemented system.

3. Tasks to perform

Below are the main tasks to be carried out in this practical work. The descriptions are generic and the examples presented are only for a better understanding of what is intended. Students must be creative and present a complete and functional integrating solution that allows them to carry out a wide variety of searches.

4.1. Analyze the Data Source (S)

The first part of the work consists of analyzing the data source and verifying where the information about the films can be found. Also, you must present a solution if there is missing data from certain films.

4.2. Define the Global Schema (G)

Define a global model for data collection. This model must be based on an XML file with the appropriate hierarchical structure to the proposed problem. That is, the student must analyze which structure of the file he considers most appropriate with regard to the level of branching and the choice of elements or attributes to store the data. The scheme to be adopted in the unified view decided by the students must always be validated using the XSD and the appropriate DTD.

4.3. Implement the Wrappers (Mappings M)

Implement the Wrappers that allow obtaining the relevant information from the data source. These wrappers must be implemented using regular expressions. The report should describe each of the wrappers in detail, indicating what information is taken by each one.
To know how to implement the Wrappers, you must analyze the structure of the HTML page where you will look for the information.
Use the ***httpRequest*** function given in the practical classes to access the pages and save them to disk.
The number and structure of the wrappers depends on the form and amount of information you want to find and must be analyzed by the students.

4.4. GENERATE / MANIPULATE XML FILE: ADD, EDIT AND DELETE DATA

After the wrappers are implemented, the data must be saved in an XML file using the chosen model.
It should be possible

- Add a new movie, wrapping data from web, as long as it does not exist in the XML file.
- If the file does not yet exist, it must be created with the insertion of the 1st film.
- Delete a movie (use the movie's name as a search word).
- Edit / change some attributes of the XML file (year, director, country, ...)

4.5. VALIDATE THE XML FILE

The XML file must be validated using the implemented XSD / DTD.
This task must be done using the JDOM2 API given in the practical classes.

4.6. SEARCH INFO USING XPATH

Allow the user to perform different searches on the XML file:

- Search for a movie title and show relevant information
- Search for films by a given director
- Search for films with the participation of a given actor / actors
- Search for movies with a duration between a given interval
- Search for films from a given country
- (other surveys proposed by students will have an additional quote)

4.7 GENERATE OUTPUTS (XSLT/XQUERY)

The program must enable the user to generate result files. These files must be transformations of the XML file of the global view.
Three mandatory transformations:

- Generate HTML file of photos of the films
- Generate XML file that shows the listing of the films of a given director
- Generate TXT file that shows the films of a given country.

Students must propose at least three additional transformations. Must implement transformations using XSLT / XQuery

4.8. GUI INTERFACE

The application must have a friendly and intuitive interface, providing the user with a set of options, for example, the following structure is suggested:

- General options
  - View the contents of the XML file
  - Validate data model (DTD and XSD)
  - Exit the application
- Change XML model data (always validate the model in each option)
  - Delete a movie from the file (use title as search word)
  - Add a movie that does not exist in the file
    - ask for the title and use the Wrappers to get the data from the web
  - Change some attributes of a film (year, director, country, ...)
- Perform XPATH Searches
  - ...
- Generate Outputs
  - ...

4. Rules and important information

The work must be carried out individually or in groups of two students.

The work values 6 points (in 20) and a minimum of 35% is required for approval in the Course.

The final work must be submitted by **June 7, 2020 at 23h55 GMT**

The delivery of the work must be done using the Moodle platform. A compressed file must be submitted whose name must contain the identification of the elements of the working group:
For example: **a22222_JohnSmith_a33333_AnnaFoster_P1.zip**

The file must contain

- the Java project with the implementation of the application
- all DTD, XSD, XSLT, XQuery, etc. files that have been implemented.
- The project report

The works will be subject to **mandatory defense** on a date to be defined.

5. Evaluation Criteria

The work is worth 6 points in the final grade of the course.
It will be evaluated according to the following criteria:

- Quality and correctness in the implementation of the requested tasks
- Program functionality
- Originality and diversification of the contents covered, namely the extra features
- Justification of the options taken
- Quality of the report delivered

Bom trabalho!
©2020 Anabela Simões