Trabalho Prático

Integração de Dados com XML – ESCRITORES

***Nota prévia**: O enunciado é propositadamente vago, genérico e incompleto em alguns pontos. O que se pretende é que os alunos avaliem as várias opções existentes e escolham a que considerarem mais apropriada para cada uma das situações com que se depararem. Todas as escolhas devem ser referidas e devidamente justificadas no relatório a entregar.*

1. OBJETIVOS

Com este trabalho pretende-se criar um programa em Java composto por vários Wrappers que obtenham dados de fontes heterogéneas, distribuídas e autónomas e possibilitem ao utilizador a visualização dos dados de forma integrada.
O utilizador terá ainda a possibilidade de fazer pesquisas, acrescentar dados que respeitem os esquemas adotados e gerar ficheiros com informação selecionada.
Para a realização deste trabalho deve usar a Linguagem Java, Expressões regulares e os APIs JDOM2 e SAXON estudados nas aulas práticas.

2. RESULTADOS DA APRENDIZAGEM

Com este trabalho prático pretende-se que se adquiram as seguintes competências:

- Saber analisar uma situação típica de Integração de Dados e apresentar propostas válidas para um modelo de integração funcional, eficaz e correto;
- Capacidade de criação e manipulação de XML
- Utilização de expressões regulares
- Capacidade de realização de pesquisa de informação em ficheiros XML usando XPath e/ou XQuery
- Capacidade de efetuar transformações de ficheiros XML usando XSLT e/ou XQuery
- Capacidade de efetuar validação de ficheiros XML usando DTD e/ou XSD

3. DESCRIÇÃO DO TRABALHO

O objetivo do trabalho é criar uma aplicação de integração de dados que apresente uma visão unificada de informações relativas a escritores famosos.
A informação deverá ser extraída dos dois sites a seguir apresentados, tratada e integrada em ficheiro(s) XML.
Fontes de dados
• https://pt.wikipedia.org/wiki/
• https://www.wook.pt/

Nota: O **nome do escritor** é dado pelo utilizador

Podem ser usadas outras fontes de dados, mas a sua utilização deve ser devidamente justificada. Se a estrutura das fontes de dados usadas não for equivalente às sugeridas acima, poderão existir penalizações na nota final.

- As duas fontes de dados **S1** e **S2** são heterogéneas, autónomas e distribuídas e contêm informação relevante sobre diversos escritores e suas obras.

- O objetivo do trabalho prático consiste em efetuar **integração de dados** provenientes destas fontes de dados e construir um modelo global **G** composto por dois ficheiros XML que agreguem a informação de forma organizada e coerente.

- Ficheiro **escritores.xml** contendo a seguinte informação para cada escritor. Esta informação, à exceção do identificador, deve ser obtida do site *wikipedia*.
    - Identificador único (gerado pelo aluno), nome, data de nascimento, data de morte (se for o caso), nacionalidade, fotografia, género literário onde se enquadra, ocupações, prémios e outra informação que considere relevante.

- Ficheiro **obras.xml** deve conter, para cada escritor do ficheiro anterior, um conjunto de obras desse autor (2 a 5 obras) publicadas. Esta informação deve ser obtida do site *wook.pt*
    - Isbn, código do autor (permite a ligação dos dois ficheiros), nome do autor, titulo, editora, preço, foto de capa.

O esquema a adotar na vista unificada deve ser decidido pelos alunos e validado usando o XSD e o DTD apropriado.

Depois de realizado o processo de integração dos dados, o utilizador poderá fazer pesquisas sobre a vista unificada.

4. TAREFAS A REALIZAR

Encontram-se em seguida as tarefas principais a desenvolver neste trabalho prático. As descrições são genéricas e os exemplos apresentados servem apenas para uma melhor compreensão do que é pretendido. Os alunos devem ser criativos e apresentar uma solução integradora completa e funcional que permita efetuar uma grande diversidade de pesquisas.

4.1. ANALISAR AS FONTES DE DADOS (S)

A primeira parte do trabalho consiste em analisar as fontes de dados e verificar onde pode ser encontrada a informação sobre os escritores e respetivas obras.

Todas as situações de exceção devem avaliadas e as decisões tomadas devem ser justificadas no relatório. Por exemplo:

- Caso encontre informação duplicada nas várias fontes
- Se um autor não for não for encontrado
- Se algum dos atributos não existir
- ...

4.2. DEFINIR O ESQUEMA GLOBAL (G)

Defina um modelo global para a recolha dos dados. Este modelo deve ser baseado em dois ficheiros XML com a estrutura hierárquica adequada ao problema proposto. Isto é, o aluno deve analisar qual a estrutura do ficheiro que considera mais adequada no que se refere ao nível de ramificação e à escolha de elementos ou atributos para guardar os dados. O esquema a adotar na vista unificada decidido pelos alunos deve ser sempre validado usando o XSD e o DTD apropriado.

Os dois ficheiros devem poder ser relacionados usando um atributo de ligação (identificador e nome do autor).

4.3. IMPLEMENTAR WRAPPERS (MAPEAMENTOS M)

Implementar os *Wrappers* que permitam obter a informação relevante de cada fonte de dados. Estes *Wrappers* devem ser implementados usando expressões regulares. No relatório deve ser descrito detalhadamente cada um dos wrappers, indicando que informação é retirada por cada um deles da fonte de dados em que cada um opera. Para cada atributo a encontrar, deve(m) ser selecionada(s) a(s) fonte(s) de dado(s) relevante(s). No caso de encontrar inconsistências ou conflitos os alunos terão de propor uma solução.

Para saber como implementar os Wrappers deve analisar a fonte das páginas HTML onde vai procurar a informação.
Use a função *httpRequest* dada nas aulas práticas para aceder às páginas e gravá-las em disco.
O número e a estrutura dos wrappers depende da forma e da quantidade de informação que se quer encontrar e deve ser analisada pelos estudantes.
A palavra de pesquisa introduzida pelo utilizador é sempre o nome do escritor. No Moodle encontra-se um ficheiro de texto com alguns nomes que podem depois ser usados para testar o programa.

4.4. GERAR / MANIPULAR FICHEIRO XML: ACRESCENTAR, EDITAR E ELIMINAR DADOS

Depois de implementados os *wrappers*, os dados devem ser guardados nos dois ficheiros XML (**escritores.xml** e **obras.xml**) usando o modelo escolhido. Deverá ser possível

- Adicionar um novo escritor e respetivas obras (sem repetições) nos ficheiros XML.
- Se os ficheiros não existirem ainda, devem ser criados com a inserção do primeiro autor e obras selecionadas.
- Eliminar um escritor (usar nome do escritor como palavra de pesquisa) e as respetivas obras do segundo ficheiro também devem ser eliminadas.
- Editar/alterar alguns atributos do ficheiro XML de escritores. Por exemplo, alterar a data de nascimento, alterar a nacionalidade, acrescentar/eliminar um prémio.

4.5. VALIDAR O MODELO G

Os ficheiros do modelo **G** devem ser validados usando os XSD/DTD escolhidos.

Esta tarefa deve ser feita usando o API JDOM2 dado nas aulas práticas.

4.6. FAZER PESQUISAS XPATH

Permitir ao utilizador efetuar diferentes pesquisas sobre o ficheiro XML:

- Pesquisar pelo nome do autor e mostrar a informação relevante (...)
- Pesquisar autores com uma nacionalidade especifica
- Pesquisar as obras de um determinado autor
- Qual o escritor mais premiado?
- Quais os livros publicados por uma determinada editora, com preço acima de um dado valor (pedir dados ao utilizador)
- (outras pesquisas propostas pelos alunos terão cotação adicional)

4.7 GERAR FICHEIROS DE OUTPUT (XSLT/XQUERY)

O programa deve possibilitar ao utilizador gerar ficheiros de resultados. Estes ficheiros devem ser transformações do ficheiro XML da vista global.

Estas quatro transformações são **obrigatórias**:

- Gerar um ficheiro HTML com uma tabela contendo duas colunas: nome dos autores, fotografias dos autores;
- Gerar ficheiro TXT que mostre a listagem de todos os autores;
- Gerar um ficheiro XML com informação dos 5 livros mais caros;
- Indicar o nome de um autor e gerar ficheiro HTML de fotos das capas dos livros desse autor
- Gerar um ficheiro XML que faça a junção da informação dos escritores e suas obras

- Os alunos devem propor no mínimo **mais três** transformações adicionais. Devem implementar as transformações usando as duas tecnologias dadas nas aulas - XSLT e XQuery – optando pela que for mais adequada em cada situação.

4.8. INTERFACE GRÁFICO

A aplicação deve ter uma interface amigável e intuitiva, disponibilizando ao utilizador um conjunto de opções, por exemplo, sugere-se a seguinte estrutura:

- Opções gerais
  - Ver conteúdo dos ficheiros XML
  - Validar modelo de dados (DTD e XSD)
  - Sair da aplicação

- Alterar dados do modelo XML (efetue sempre a validação do modelo em cada uma das opções)
  - Eliminar um autor do ficheiro **escritores.xml**
    - usar nome como palavra de pesquisa
    - eliminar também as obras desse autor do ficheiro **obras.xml**
  - Acrescentar um escritor que não exista no ficheiro
    - Acrescentar também um máximo de 5 obras desse autor no ficheiro **obras.xml**
  - Alterar alguns atributos de um escritor
    - Por exemplo, alterar a data de nascimento, alterar a nacionalidade, acrescentar/eliminar um prémio, ...

- Efetuar Pesquisas XPATH
  - ...

- Gerar Outputs
  - ...

5. NORMAS PARA REALIZAÇÃO DO TRABALHO

O trabalho deverá ser realizado **individualmente ou em grupos de dois alunos**.

O trabalho vale 6 valores e é necessário um mínimo de 35% para aprovação na Unidade Curricular.

O trabalho final deve ser entregue até **11 de Junho de 2023** às 23h55 GMT.

>>>>>> DATA ÚNICA DE ENTREGA PARA TODAS AS ÉPOCAS DE EXAME <<<<<<

A entrega dos trabalhos deverá ser feita usando a plataforma Moodle. Deve ser submetido um ficheiro compactado cujo nome deve conter a identificação dos elementos do grupo de trabalho:

Por exemplo: **a22222_AnaMatos_a33333_RuiMelo_P1.zip**

O ficheiro deve conter o projeto Java com a implementação da aplicação e todos os ficheiros DTD, XSD, XSLT, XQuery, etc que foram implementados.

Os trabalhos serão sujeitos a **defesa obrigatória** nas aulas das semanas 12 a 23 de Junho.

6. CRITÉRIOS DE AVALIAÇÃO

O trabalho vale **6 valores** na nota final da Unidade Curricular. Para aprovação na UC é necessário ter um mínimo de 35% neste trabalho.

O trabalho será avaliado segundo os seguintes critérios:

- Qualidade e correção na implementação das tarefas solicitadas
- Funcionalidade do programa
- Originalidade e diversificação dos conteúdos abordados, nomeadamente as funcionalidades extras
- Justificação das opções tomadas
- Qualidade do relatório entregue
- Qualidade da defesa

Bom trabalho!