

Trabalho Prático
Integração de Dados com XML

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 países.

A informação deverá ser extraída dos dois sites a seguir apresentados, tratada e integrada em ficheiro(s) XML.

As fontes de dados a usar são:

- S1 - https://en.wikipedia.org/wiki/nome_país
(por exemplo: <https://en.wikipedia.org/wiki/Spain>)
- S2 - https://www.countryreports.org/country/nome_país.htm
(por exemplo: <https://www.countryreports.org/country/spain.htm>)

Nota: **nome_país** é dado pelo utilizador. Para simplificação da utilização dos dois sites, sugere-se o use o nome dos países em inglês

Outros links da wikipedia que pode explorar para obter os dados:

- Países do continente Europeu: <https://en.wikipedia.org/wiki/Europe>
- Países do continente Africano: <https://en.wikipedia.org/wiki/Africa>
- semelhante para os outros continentes
- Lista de países válidos: https://en.wikipedia.org/wiki/List_of_sovereign_states

Podem ser usadas outras fontes de dados, mas a sua utilização deve ser devidamente justificada e mencionada no relatório. 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 heterogêneas, autónomas e distribuídas e contêm informação relevante sobre diversos países.
- 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 ficheiro XML que agregue a informação de forma organizada e coerente.
 - O ficheiro **países.xml** deve conter (pelo menos) a seguinte informação: código ISO do país, nome do país, continente a que pertence, nome do Presidente da República, link para a imagem da bandeira do país.
 - O ficheiro **factos.xml** deve conter (pelo menos) a seguinte informação: código ISO do país, nome do país, capital, moeda, população, área em km², crescimento da população, cidades mais populosas , domínio internet, lista de idiomas oficiais, lista de países vizinhos

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 países.

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 país 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 ISO).

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 país. No Moodle encontra-se um ficheiro de texto com alguns nomes que podem ser usados para testar o programa. Use outros países adicionalmente!!!

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 (**países.xml** e **factos.xml**) usando o modelo escolhido. Deverá ser possível

- Adicionar um novo país e respetivos factos (sem repetições) nos ficheiros XML.
- Se os ficheiros não existirem ainda, devem ser criados com a inserção do primeiro país e factos respetivos.
- Eliminar um país (usar nome do país como palavra de pesquisa) e os respetivos factos do segundo ficheiro também devem ser eliminados.
- Editar/alterar alguns atributos do ficheiro XML de países. Por exemplo, alterar capital, alterar população, alterar/adicionar um idioma, etc).

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 país e mostrar a informação relevante (selecione a informação mais importante)
- Pesquisar países de um determinado continente
- Pesquisar os factos de um determinado país
- Qual o país mais populoso?
- Quais os países de um determinado continente, com população acima de um dado valor (pedir estes 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 países, fotografias das bandeiras dos países;
- Gerar ficheiro TXT que mostre a listagem de todos os países;
- Gerar um ficheiro XML com informação dos 5 países mais populosos;
- Indicar o nome de um continente e gerar ficheiro HTML de fotos das bandeiras dos países desse continente
- Gerar um ficheiro XML que faça a junção da informação de um país e dos respetivos factos.
- 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 país do ficheiro **países.xml**
 - usar nome do país como palavra de pesquisa
 - eliminar automaticamente os factos desse país do ficheiro **factos.xml**
 - Acrescentar um país que não exista no ficheiro
 - Acrescentar automaticamente os factos desse país do ficheiro **factos.xml**
 - Alterar alguns atributos de um país
 - Por exemplo, alterar a capital, alterar a população, acrescentar/eliminar um idioma, ...
- Efetuar Pesquisas XPATH
 - ...
- Gerar Outputs
 - ...

5. NORMAS PARA REALIZAÇÃO DO TRABALHO

O trabalho deverá ser realizado **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 **19 de Maio de 2024** 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 ZIP deve conter:

- o relatório
- o projeto Java com a implementação da aplicação
- todos os ficheiros DTD, XSD, XSLT, XQuery

Os trabalhos serão sujeitos a **defesa obrigatória** nas aulas das semanas 20 a 23 de Maio.

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!
©2024 Anabela Simões