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. OBJECTIVOS

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 adoptados e gerar ficheiros com informação selecionada.

Para a realização deste trabalho deve usar a Linguagem Java, Expressões regulares e o API JDOM2 para manipulação de XML estudado 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 correcto;
- Capacidade de criação e manipulação de XML
- Utilização de expressões regulares
- Capacidade de realização de pesquisa de informação em ficheiros XML usando XPath ou XQuery
- Capacidade de efetuar transformações de ficheiros XML
- Capacidade de efetuar validação de ficheiros XML usando DTD e/ou XSD

3. DESCRIÇÃO DO TRABALHO

O objectivo do trabalho consiste na criação de uma aplicação integradora que apresente uma vista unificada de informação relativa a escritores de várias nacionalidades, contendo informação biográfica e obras publicadas em Portugal. As fontes de dados a usar são:

- S1 - http://pt.wikipedia.org/wiki/**nome_do_escritor**
- S2 - http://www.wook.pt/

Nota: **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 objectivo 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 excepção 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 (2 a 5 obras) publicadas em Portugal. Esta informação deve ser obtida do site *wook*
    - Isbn, código do autor, titulo, editor, preço, foto de capa, ano, ...

O esquema a adoptar 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 efectuar uma grande diversidade de pesquisas.

4.1. IMPLEMENTAR WRAPPERS

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.

4.2. GERAR FICHEIROS XML

A informação obtida pelos Wrappers deve ser usada para a criação dos ficheiros XML de acordo com os modelos descrito na secção 3.

4.3. 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.

4.4. DEFINIR PESQUISAS XPATH

Permitir ao utilizador efectuar diferentes pesquisas:

1) Sobre os ficheiros XML criados
2) Sobre as fontes de dados originais
    - **Exemplo: obter detalhes de um escritor introduzido pelo utilizador**
        - Escritor já está no ficheiro escritores.xml? → mostrar dados biográficos e obras publicadas.
        - Escritor não está no ficheiro? → procurar nas fontes de dados usando os wrappers, mostrar ao utilizador os dados biográficos e obras publicadas e adicionar a nova informação aos ficheiros XML, respeitando o modelo G. Se o escritor não for encontrado, informar o utilizador do insucesso da pesquisa.

- Os passos 1) e 2) devem ser aplicados a diferentes tipos de pesquisas:
- Exemplos de pesquisas:
    - **Procurar escritores de uma nacionalidade especifica**
    - **Procurar escritores de um determinado género literário**
    - **Introduzir um titulo/isbn de um livro e obter os dados do escritor respectivo**
    - **Procurar os livros de um determinado escritor**
    - **Top 5 dos livros mais baratos do ficheiro**
    - **Qual o escritor mais premiado?**
    - **Efectuar pesquisas que combinem dois ou mais atributos**
    - ...

As pesquisas devem ser definidas pelos alunos, devendo ser variadas e versáteis com combinação de um diferente número de atributos.
A escolha dos atributos de pesquisa deve ser escolhido pelo utilizador usando Interface Java.
Os resultados devem ser apresentados de forma atrativa e organizada usando Interfaces Java.
As pesquisas devem ser implementadas usando XPath.

4.5. EDITAR E ELIMINAR INFORMAÇÃO

O interface deve possibilitar ao utilizador editar um determinado atributo dos ficheiros XML. Por exemplo, alterar a data de nascimento, alterar a nacionalidade, acrescentar/eliminar um prémio, acrescentar/eliminar uma obra, etc. O utilizador também pode eliminar um determinado autor do ficheiro **escritores.xml**
A informação dos dois ficheiros deve estar coerente e relacionada usando um campo à escolha do aluno. A remoção de um autor do ficheiro **escritores.xml** deve ter como consequência a remoção de todas as suas obras do ficheiro **obras.xml**. Após as alterações, o modelo deve ser validado, para garantir a sua integridade.

4.6 GERAR FICHEIROS DE OUTPUT

O programa deve possibilitar ao utilizador gerar ficheiros de resultados. Estes ficheiros devem ser transformações dos ficheiros XML da vista global.
Duas transformações **obrigatórias**:

- Gerar um ficheiro HTML contendo as fotografias de todos os escritores listadas no ficheiro XML;
- Gerar um novo ficheiro XML que junte atributos dos dois ficheiros: por exemplo, juntar num mesmo ficheiro XML o nome do autor e a lista dos títulos publicados por esse escritor.

Os alunos devem propor no mínimo **mais três** transformações adicionais. Devem implementar as transformações usando XSLT / XQuery

5. NORMAS PARA REALIZAÇÃO DO TRABALHO

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

A indicação dos elementos do grupo de trabalho é obrigatória e deve ser feita por email (abs@isec.pt) até dia 10 de Abril. Quem não fornecer estes dados não poderá defender o trabalho.

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

O trabalho final deve ser entregue até **10 de junho de 2018** às 23h55 GMT. Os trabalhos serão sujeitos a **defesa obrigatória** na aula prática da semana 11-14 de junho. Neste dia, deverá ser entregue o relatório em papel.

6. CRITÉRIOS DE AVALIAÇÃO

O trabalho vale **6 valores** na nota final da Unidade Curricular.
Será avaliado segundo os seguintes critérios:

- Qualidade e correcção 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!