Integração de Dados

Licenciatura em Engenharia Informática: 2º ano - 2º semestre 2020/2021

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

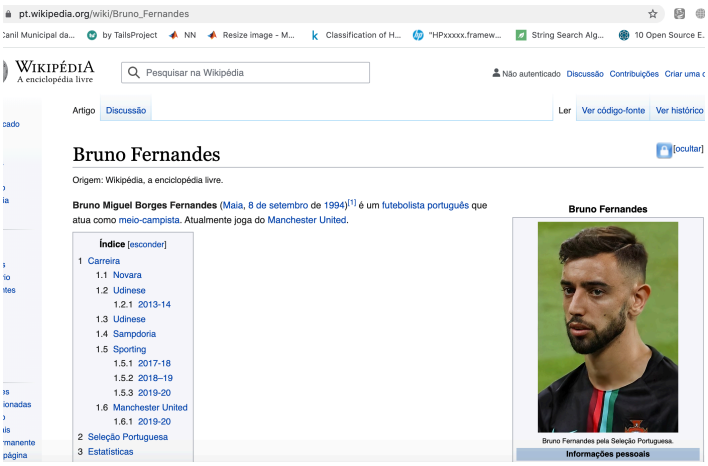
O objetivo do trabalho consiste na criação de uma aplicação integradora que apresente uma vista unificada de informação relativa a jogadores de futebol, proveniente de diferentes páginas da Internet:

- http://zerozero.pt
- https://pt.wikipedia.org/wiki/
- https://www.transfermarkt.pt/

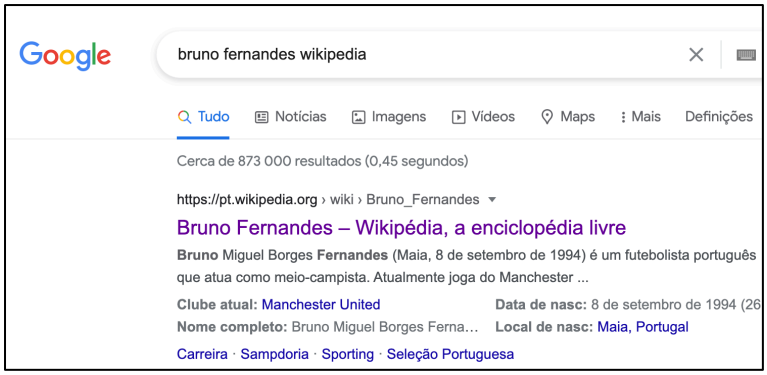
Utilize a função *HttpRequest* disponibilizada no Moodle, usando o link geral e a palavra de pesquisa (nome do jogador) em cada um dos sites. Por exemplo:

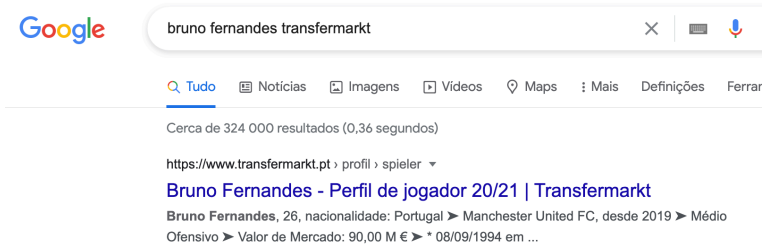
```
HttpRequestFunctions.httpRequest1("https://pt.wikipedia.org/wiki/", "Bruno Fernandes", "jogador.txt");
```

Acede corretamente ao site do jogador:



Em alguns jogadores podem ocorrer situações de não haver um acesso direto ao site pessoal, e para desambiguação aconselha-se o uso do Google para encontrar o link direto para cada site. Nas imagens seguintes mostra-se o exemplo do jogador Bruno Fernandes, onde usando pesquisas Google será possível através de ERs obter os links diretos para as diferentes páginas onde poderá posteriormente obter a informação desejada.





As fontes de dados são heterogéneas, autónomas e distribuídas e contêm informação relevante sobre o tema do trabalho.

O objetivo do trabalho prático consiste em efetuar **integração de dados** provenientes destas várias fontes de dados e construir um modelo global de dados usando XML. Este modelo de dados deve ser constituído por **um ficheiro XML** onde toda a informação pesquisada seja organizada em elementos/atributos, usando a hierarquia decidida pelos alunos como a mais correta para executar as tarefas propostas.

A informação que se pretende guardar no ficheiro é a seguinte (os alunos devem incluir informação adicional que achem relevante):

- nome completo do jogador
- nome pelo qual é mais conhecido
- fotografia (link)
- data de nascimento, idade
- nacionalidade
- altura
- peso
- pé preferencial
- posições onde joga
- clube atual
- clubes anteriores
- seleção nacional (se for o caso)
- prémios ganhos (taça, dat, etc)
- estado atual: ativo, reformado
- valor do contrato
- empresário
- ranking dado pelo site transfer markt
- ....

Os alunos devem analisar as diferentes fontes de dados e apresentar um estudo das mesmas, decidindo e justificando onde vão retirar a informação. Obrigatoriamente, Ddevem ser usadas **pelo menos duas** das fontes de dados fornecidas.

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.

No moodle encontra-se uma lista de jogadores que podem servir de teste da aplicação. Depois de terminada a aplicação, os alunos devem testar com outros nomes e avaliar a funcionalidade do sistema implementado.

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 jogadores.
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 jogador não for encontrado
- Se algum dos atributos pedido não existir para alguns jogadores
- ...

4.2. DEFINIR O ESQUEMA GLOBAL (G)

Defina um modelo global para a recolha dos dados. Este modelo deve ser baseado num ficheiro 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.

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 estrutura 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.

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

Depois de implementados os *wrappers*, os dados devem ser guardados num ficheiro XML usando o modelo escolhido. Deverá ser possível

- Adicionar um novo jogador, desde que este não exista no ficheiro XML.
- Se o ficheiro não existir ainda, deve ser criado com a inserção do 1º jogador.
- Eliminar um jogador (usar nome como palavra de pesquisa).
- Editar/alterar alguns atributos do ficheiro XML (idade, nacionalidade, clube atual, ...)

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 de um jogador (pode ser parcial) e mostrar a informação relevante
- Pesquisar jogadores de um dado clube (atual)
- Pesquisar jogadores com uma determinada nacionalidade
- Pesquisar jogadores que joguem numa dada posição
- Pesquisar jogadores que não estejam no ativo (reformados, ou com outras funções)
- (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.
quatro transformações **obrigatórias**:

- Gerar ficheiro HTML de fotos dos jogadores do ficheiro
- Gerar ficheiro XML que mostre a listagem dos jogadores de um dado clube
- Gerar ficheiro TXT que mostre os nomes dos jogadores com uma determinada nacionalidade.
- Gerar um ficheiro XML com top 5 dos jogadores com contratos mais valiosos

Os alunos devem propor no mínimo **mais três** transformações adicionais. Devem implementar as transformações usando as duas tecnologias dadas nas aulas: XSLT e XQuery

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 do ficheiro 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 jogador do ficheiro (usar nome como palavra de pesquisa)
  - Acrescentar um jogador que não exista no ficheiro
    - pedir o nome e usar os Wrappers para obter os dados da web
  - Alterar alguns atributos de um jogador (idade, clube, nacionalidade, ...)
- 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 **06 de Junho de 2021** 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 7 a 18 de Junho.

6. CRITÉRIOS DE AVALIAÇÃO

O trabalho vale **6 valores** na nota final da Unidade Curricular.
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!