Dados Semiestruturados – XML

Exercício 1

Descrição Geral

Neste exercício, você desenvolverá um programa em JavaScript para manipular documentos XML que são *feeds* RSS. Você precisará exibir partes do conteúdo do documento em uma página em HTML criada pelo seu programa. O seu programa JavaScript deverá ficar embutido em um arquivo HTML, para que ele possa ser executado diretamente em um *browser*.

De acordo com a Wikipédia, "*os feeds RSS oferecem conteúdo Web ou resumos de conteúdo juntamente com os links para as versões completas deste conteúdo e outros metadados. Esta informação é entregue como um arquivo XML chamado "RSS feed", "webfeed", "Atom" ou ainda canal RSS.*" – https://pt.wikipedia.org/wiki/RSS.
Muitos sites que mudam o seu conteúdo regularmente, como os de notícias e os *blogs*, fornecem canais (*feeds*) RSS para o seu conteúdo. E aplicações conhecidas como *feed readers* permitem que um usuário cadastre nelas URL de canais RSS de seu interesse e exibem os dados desses canais (atualizados periodicamente) em uma interface amigável. Estas aplicações são tipicamente construídas como programas independentes (ex.: aplicativos de celular) ou como extensões de navegadores ou programas de correio eletrônico.
A sigla RSS, na versão atual, se refere a *Really Simple Syndication*. Para saber mais detalhes sobre o formato RSS, acesse: http://www.w3schools.com/xml/xml_rss.asp .

Usaremos como entrada para o programa deste exercício documentos XML de canais RSS do jornal New York Times (NYT). Na página http://www.nytimes.com/services/xml/rss/index.html vocês encontram uma lista dos canais RSS disponibilizados pelo NYT e suas respectivas URLs. Um exemplo de canal dessa lista do NYT é o de "Technology" e "World":
http://rss.nytimes.com/services/xml/rss/nyt/Technology.xml
http://rss.nytimes.com/services/xml/rss/nyt/World.xml

Para ver o conteúdo XML de um canal, abra a URL dele no *browser,* depois clique com o botão esquerdo do *mouse* sobre a página aberta (para abrir o menu de contexto) e selecione a opção "Ver código fonte da página". Obs.: O nome dessa funcionalidade e a forma de acessá-la pode variar de um *browser* para outro.

O seu programa deverá abrir e validar o documento XML de um canal RSS e depois listar em uma página HTML, na forma de links, as categorias de notícias contidas no documento. Quando o usuário clicar num dos links de categoria mostrados na página, o programa deve exibir na página uma lista com os títulos e links para as notícias extraídas do documento que estão na categoria do link clicado.

Descrição mais detalhada do que deve ser feito:

1. Inicialmente, o seu programa deve mostrar uma página web com um campo onde o usuário possa digitar ou selecionar o nome e caminho completo do documento XML que ele quer fornecer como entrada para o programa, e um botão para ele fazer a envio do valor.

2. Se o usuário não preencher um valor para o campo nome do arquivo e clicar no botão enviar, uma mensagem de erro apropriada deve ser exibida e seu programa deve permanecer na página inicial.

   Importante: neste exercício, você sempre vai lidar com **arquivos XML armazenados localmente, na mesma pasta do programa**. Dois documentos RSS de exemplo, chamados **nyt_world_rss.xml** e **nyt_technology_rss.xml** (obtido do site do NYT) estão disponíveis junto com este enunciado no Paca.
   O acesso via JavaScript a arquivos XML localizados em outras pastas e servidores tem restrições por motivos de segurança. Para saber mais sobre esse assunto, veja:
   https://www.google.com/about/appsecurity/learning/xss/

3. Quando um usuário fornece um valor para o campo de nome do arquivo e clica no botão enviar, o seu programa deve verificar se o valor fornecido corresponde ao caminho de um documento RSS válido. Para isso, você deve verificar se o conteúdo dele está em conformidade com o esquema em XML Schema definido no arquivo **rss-2_0.xsd** (disponibilizado junto com este enunciado no Paca). Caso o conteúdo não seja válido segundo esse esquema, uma mensagem de erro apropriada deve ser exibida.

4. Uma vez que o documento XML passe com sucesso pela validação, o programa deverá fazer uma consulta para recuperar do documento todas as categorias de notícias contidas nele. Observe que, em um arquivo RSS, cada notícia (elemento **item**) pode ser classificada em várias categorias (subelementos **category**). Cada categoria recuperada deve ser listada numa página web como um link. Cada categoria deve aparecer na listagem uma só vez (ainda que apareça no documento XML inúmeras vezes).

5. Quando um usuário clicar em um dos links de categoria criados no item anterior, o programa deverá exibir uma tabela com o título e o link de todas as notícias contidas no documento classificadas na categoria do link clicado.

Programa de Exemplo

Na pasta "Exemplo", disponibilizada junto com este enunciado, vocês encontram um exemplo de programa em JavaScript/HTML (arquivo "exemplo_programa.html") que abre um documento em XML (arquivo "catalogo_cds.xml"), verifica se ele está em conformidade com um dado esquema em XML Schema (arquivo "catalogo_cds.xml"), faz uma consulta para recuperar alguns elementos do documento e depois exibe-os em uma tabela na página web. Esse exemplo já contém (quase) tudo de que vocês precisam para resolver o exercício.
Para executar o programa, basta abri-lo no *browser* **Firefox**. Atenção: para que o programa funcione no Chrome, é preciso abrir o *browser* com uma configuração especial, como mostrado a seguir:

```
$ chrome -allow-file-access-from-files
```

Referências Úteis

- HTML Forms:
  http://www.w3schools.com/html/html_forms.asp

- JavaScript:
  http://www.w3schools.com/js/default.asp
  https://developer.mozilla.org/en-US/docs/Web/JavaScript

- Documentação da API Web (Mozilla)
  https://developer.mozilla.org/en-US/docs/Web/Reference/API

- Mais detalhes sobre o XMLHttpRequest:
  https://developer.mozilla.org/pt-BR/docs/Web/API/XMLHttpRequest

- O esquema em XML Schema para arquivos RSS 2.0 que usamos neste exercício é uma adaptação do esquema criado por Jorgen Thelin, que está protegido pela licença *Microsoft Public License* (Ms-PL). O esquema original está disponível publicamente em: http://rss2schema.codeplex.com/releases/view/18981 .