

A LINGUAGEM XML

Engenharia de Dados e Conhecimento
2019/2020

O que é? (i)

- XML é um dos formatos mais populares da indústria utilizados quer na publicação de documentos quer no desenvolvimento de aplicações web.
- É uma solução extensível e elegante rapidamente incorporada nos documentos de próxima geração e em aplicações web.
- Por causa da sua semelhança com HTML, o XML é ideal como forma de transferir dados através da Internet - por exemplo, normalmente não é bloqueado por *firewalls* (porto 80), enquanto os aplicativos personalizados, usando portos próprios para o transporte de dados são muitas vezes inibidos.

O que é? (ii)

- O XML é um subconjunto do SGML. A SGML (*Standard Generalized Markup Language*) definiu um conjunto de especificações para a partilha de dados entre aplicações de grande porte.
- A SGML é uma linguagem extremamente complexa, e embora tenha sido vista como potencialmente útil, poucas aplicações realmente se aproveitaram disso na prática.
 - Isto levou ao surgimento da piada SGML = “Sounds good – maybe later”
- No entanto foi esta linguagem que levou ao desenvolvimento de duas linguagens muito utilizadas:
 - a HTML, para a construção de páginas web;
 - a XML, para compartilhar dados.


O que é? (iii)

- O XML não é uma linguagem mas uma "meta-linguagem".
- Isso significa que é uma **ferramenta** que permite criar linguagens específicas para descrever dados.
 - Por exemplo: uma empresa de azulejos pode criar uma linguagem própria para descrever os seus produtos com campos como "Dimensão", "Material" e "Lavável".
- O único requisito para o seu bom funcionamento é todos os seus utilizadores conhecerem e aplicarem bem as regras utilizadas na sua criação.

O que é? (iv)

- "XML" é frequentemente utilizado como “guarda chuva” para várias tecnologias relacionadas.
- O conjunto de tecnologias relacionadas inclui:
 - XML Schema - usado para definir a estrutura dos dados.
 - XSLT (Transformações) - utilizadas para pegar num conjunto de dados baseados em XML e transformá-lo noutra coisa, como uma página web.
 - XPath - usado para navegar através de um documento XML e selecionar dados específicos.

O que é? (v)

- O XML é usado por muitas outras tecnologias:
 - os *datasets* em ADO NET são baseados em XML;
 - o formato de armazenamento de ficheiros do Microsoft Office é o Open Office XML (notar o "x" nas extensões dos arquivos a partir do Office 2007 e seguintes, como "*.docx", "*.pptx", "*.xlsx", etc.);
 - aplicações que usam arquivos de configuração baseados em XML para armazenar as preferências do utilizador e opções.
 - muitas páginas com um símbolo como este () que indica que está a "assinar" um *feed* de informação. O conteúdo é fornecido em XML e por isso é tão fácil chegar da Internet ao PC, e ser apresentado da forma que lhe convier.

Estrutura de um documento XML (i)

- Fisicamente, um documento é composto por unidades chamadas "**entidades**".
- Uma entidade pode referenciar outras entidades, fazendo com que estas sejam incluídas no documento.
- Um documento XML é uma entidade finita. O tamanho finito é obtido através de marcadores iniciais e finais, os quais delimitam a entidade e o seu conteúdo.

Estrutura de um documento XML (ii)

- Logicamente, o documento é composto de declarações, elementos, atributos, comentários e instruções de processamento.
 - Todos são indicados no documento através da utilização de marcadores.
- Dois documentos XML podem diferir na estrutura física (marcadores + dados), mas podem possuir uma estrutura lógica igual (marcadores).
 - Uma das formas de verificar se dois documentos XML fisicamente diferentes possuem a mesma estrutura lógica é através da análise do subconjunto de dados nele contido e da sintaxe utilizada para expressar aquele subconjunto.

Estrutura de um documento XML (iii)

- As estruturas lógica e física têm de se encaixar convenientemente de modo a ter um documento “bem formado”.
- Num documento “bem formado” os seus elementos (nós) ligam-se de modo a formar uma estrutura em árvore.

Estruturas físicas

- Um documento XML começa com a declaração abaixo e especifica a versão do XML com a qual está conforme:

```
<?xml version="1.0"?>
```

- A declaração do documento XML **pode** também conter um atributo que identifique o conjunto de caracteres contido no documento (*encoding*).

```
<?xml version="1.0" encoding="ISO-8859-1"?>
```

```
<?xml version="1.0" encoding="UTF-8"?>
```

Estruturas físicas

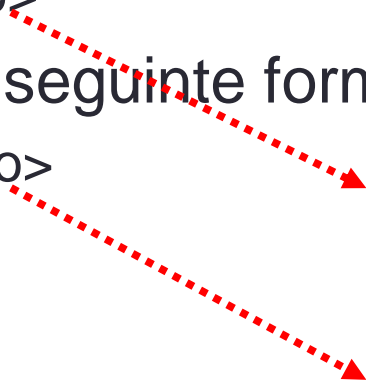
- “Uma entidade pode referenciar outras entidades”

Exemplo:

```
<?xml version="1.0" encoding="UTF-8"?>
<collection xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance
  xsi:schemaLocation="http://www.bn.pt/standards/metadata/marcxml/1.0/
    http://xml.bn.pt/schemas/Unimarc-1.0.xsd"
  xmlns="http://www.bn.pt/standards/metadata/marcxml/1.0/">
...
</collection>
```

Estruturas lógicas - Elementos

- São os blocos lógicos em que um documento pode ser decomposto
- Um elemento tem sempre a seguinte estrutura:
 - <nome_do_elemento>
- E deve terminar da seguinte forma:
 - </nome_do_elemento>




The diagram consists of two red dotted arrows. The first arrow originates from the opening tag '<nome_do_elemento>' in the list and points to the opening tag '<Aluno ID="12345">' in the code block. The second arrow originates from the closing tag '</nome_do_elemento>' in the list and points to the closing tag '</Aluno>' in the code block.

```
<Aluno ID="12345">  
  <Identificação>  
    <Nome>Maria da Silva</Nome>  
  </Identificação>  
</Aluno>
```

Estruturas lógicas - Atributos

- Aparecem sempre na declaração inicial do elemento



```
<Aluno ID="12345">  
  <Identificação>  
    <Nome>Maria da Silva</Nome>  
  </Identificação>  
</Aluno>
```

- Permitem qualificar o elemento a que estão associados.

Estruturas lógicas – Elementos Vazios

- Marcadores de **elemento vazio** podem ser utilizados com qualquer elemento que não precise de conter qualquer outro elemento



- Os atributos, quando existem, são colocados no marcador de abertura do elemento

Exemplo de um registo bibliográfico

formato nativo / texto

286126

(700): Zweig. Stefan

(702): Osswald, Maria Henriques

(200): Os grandes momentos da humanidade

(201): Stefan Zweig ; trad. Maria Henriques Osswald

(205): 5ª ed.

(210): Porto

(212): Livraria Civilização

(211): 1960

(215): Pag. var.

(217): 19 cm

(966): PP9904|CLP Goa

(931): 20100608

(606): Literatura

(920): n

(921): a

(922): m

(932): d

(935): k

(936): y

(937): 0

(938): ba

(999): n

(970): Gemira

(971): Lourdes

```

<?xml version="1.0" encoding="UTF-8"?>
<collection xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.bn.pt/standards/metadata/marcxml/1.0/ http://xml.bn.pt/schemas/Unimarc-1.0.xsd"
  xmlns="http://www.bn.pt/standards/metadata/marcxml/1.0/">
  <record>
    <leader>   nam       450 </leader>
    <controlfield tag="005">20100608000000.0</controlfield>
    <controlfield tag="009">286126</controlfield>
    <datafield tag="100" ind1=" " ind2=" ">
      <subfield code="a">-----k 0---y-----ba</subfield>
    </datafield>
    <datafield tag="200" ind1="1" ind2=" ">
      <subfield code="a">Os grandes momentos da humanidade</subfield>
      <subfield code="f">Stefan Zweig ; trad. Maria Henriques Osswald</subfield>
    </datafield>
    <datafield tag="205" ind1="1" ind2=" ">
      <subfield code="a">5ª ed.</subfield>
    </datafield>
    <datafield tag="210" ind1=" " ind2=" ">
      <subfield code="a">Porto</subfield>
      <subfield code="c">Livraria Civilização</subfield>
      <subfield code="d">1960</subfield>
    </datafield>
    <datafield tag="215" ind1=" " ind2=" ">
      <subfield code="a">Pag. var.</subfield>
      <subfield code="d">19 cm</subfield>
    </datafield>
    (...)
  </record>
</collection>

```

286126

(700): Zweig, Stefan

(702): Osswald, Maria Henriques

(200): Os grandes momentos da humanidade

(201): Stefan Zweig ; trad. Maria Henriques Osswald

(205): 5ª ed.

(210): Porto

(212): Livraria Civilização

(211): 1960

(215): Pag. var.

(217): 19 cm

(966): PP9904|CLP Goa

(931): 20100608

(606): Literatura

...

Dúvidas

- Mas o XML é só uma norma para escrever os nossos dados em formato texto?
- Será possível "**descrever**", e mesmo "**validar**", a estrutura de um documento XML?
- Isso é fundamental para que esta norma possa ser utilizada na troca de informação entre sistemas ...

- Numa primeira fase a resposta à segunda pergunta chamou-se DTD's - *document type definitions*.
- Actualmente a resposta a esta pergunta chama-se *XML Schema*.

Elementos vs Atributos

Pergunta de um milhão \$: Quando devem ser utilizados elementos e quando devem ser utilizados atributos?

- Elementos podem conter outros elementos, mas atributos não podem conter outros atributos ou outros elementos.
- Atributos e Elementos Simples permitem especificar:
 - Gammas de valores possíveis
 - Valores por defeito