XML & XSL

da teoria à prática

José Carlos Leite Ramalho

Setembro de 2005

Motivação 1

- Um dia de trabalho = produção de vários documentos
- Muitos têm uma forma estruturada
- Alguns até podem ser representados numa tabela: inventários, preçários, ...
- Mas, apenas 10% da informação é guardada em Bases de Dados
- Que fazer com os restantes 90%?

Os restantes 90%

- Correspondem a textos que circulam dentro das instituições
- Não se lhes pode aplicar uma metodologia relacional
- Haverá alguma maneira de contornar o problema?

- A solução recai sobre a estruturação da informação

Exemplo: uma carta

Exmo Vice-Reitor
Prof. Dr. José Viriato Eiras Capela

Devido à proximidade do prazo e ao trabalho em que ... venho, por este meio, solicitar-lhe que conceda
mais 30 dias para a submissão final da tese de mestrado a dois dos meus orientandos: Joel Vicente (Mestrado em Informática) e Luis Miguel Alves Domingues (Mestrado em Informática).
...

Com os melhores cumprimentos

Universidade do Minho, Braga,
21 de Setembro de 2005

José Carlos Leite Ramalho
(Professor Auxiliar)

Motivação 2

- Publicação Electrónica
  - Proliferação das TICs = proliferação de formatos
  - Explosão da Web veio agravar ainda mais
  - Questão: Como conseguir produzir documentos num formato neutro a partir do qual seja possível gerar todos os formatos necessários para distribuição?

Conteúdo

- Documentação Estruturada
  - Anotação
    - Procedimental
    - Descritiva
  - Linguagens de Anotação
  - Tipos de Anotação
  - Evolução das Linguagens de Anotação
    - HTML versus XML

Documentação Estruturada

- Valor de um documento = facilidade na localização, no consumo, na validação e na reutilização
- Um documento estruturado tem as seguintes vantagens:
  - Acesso
  - Validação
  - Reutilização
  - Normalização

Anotação

- “Markup” = anotação, codificação, etiquetagem
- A anotação de um texto é um meio de tornar explícita uma interpretação desse texto
- Exemplo:
  - “Está a chover.”
  - “Está a chover?”

Objectivos da Anotação

1. Dividir o documento em componentes
   - Dá organização lógica (explicitamente)
   - Dá indicações para o processamento (implicitamente)
2. Associar semântica
   - Dá interpretação (implicitamente)
   - Dá indicações para a formatação (explicitamente)

Funções da Anotação

- representar todos os caracteres de um texto
- identificar a estrutura do texto
- reduzir o texto a uma ordem linear (árvore)
- representar informação contextual
- distinguir o que é texto do que é anotação

Fases da Anotação

1. Análise da estrutura da informação (dos documentos que se pretende tratar).
2. Definição da formatação/transformação desejada para cada elemento estrutural.
3. Inserção das anotações no documento.

Pausa para pensar

- Exercício: anotar o poema
- Exercício: definir as anotações para a agenda
- Exercíco: anotar um relatório

Exercício: o poema

"Soneto Já Antigo"
(Álvaro de Campos)

Olha, Daisy: quando eu morrer tu hás-de
dizer aos meus amigos aí de Londres,
embora não o sintas, que tu escondes
a grande dor da minha morte. Irás de

Londres p'ra Iorque, onde nasceste (dizes
que eu nada que tu digas acredito),
contar áquele pobre rapazito
que me deu horas tão felizes,

embora não o saibas, que morri...
Mesmo ele, a quem eu tanto julguei amar,
nada se importará... Depois vai dar

a notícia a essa estranha Cecily
que acreditava que eu seria grande...
Raios partam a vida e quem lá ande!

(1922)

Poema: título, autor, corpo, data
Corpo: quadra, quadra, terno, terno.
Quadra: verso, verso, verso, verso
Terno: verso, verso, verso
Verso: (texto | nome)+
Nome: texto

Anotação Procedimental

```
Exmo Vice-Reitor
Prof. Dr. José Viriato Eiras Capela
.vspace
Devido à proximidade do prazo e ao trabalho em que ... venho, por este
meio, solicitar-lhe que conceda
mais 30 dias para a submissão final da tese de mestrado a dois dos meus
orientandos: Joel Vicente (Mestrado em Informática) e Luis Miguel Alves
Domingues (Mestrado em Informática).
...
.vspace
.indent 16
Com os melhores cumprimentos
.vspace
Universidade do Minho, Braga,
21 de Setembro de 2005
.center
José Carlos Leite Ramalho
(Professor Auxiliar)
```

Define qual o processamento a ser realizado em determinados pontos do documento.

Anotação Descritiva

```
<carta>
<destinatario> Exmo Vice-Reitor
Prof. Dr. José Viriato Eiras Capela </destinatario>
<corpo>
Devido à proximidade do prazo e ao trabalho em que ... venho, por este
meio, solicitar-lhe que conceda
mais 30 dias para a submissão final da tese de mestrado a dois dos meus
orientandos: Joel Vicente (Mestrado em Informática) e Luis Miguel Alves
Domingues (Mestrado em Informática).
... </corpo>

<fecho> Com os melhores cumprimentos
Universidade do Minho, Braga,
21 de Setembro de 2005
José Carlos Leite Ramalho
(Professor Auxiliar) </fecho>
</carta>
```

Utiliza etiquetas para apenas classificar as componentes do documento.

Linguagem de Anotação

- Especifica como distinguir a anotação do texto
- Especifica **que** anotações são **necessárias** e **quais** são **permitidas**
- Especifica **onde** as anotações são **necessárias** e **onde** são **permitidas**
- Define o significado da anotação

O XML tem estas características todas ... excepto a última

Perspectivas de Anotação

1. Anotação orientada ao formato
2. Anotação orientada à estrutura
3. Anotação orientada ao conteúdo

- Objectivo: Anotação Equilibrada

Anotação orientada ao formato

```
<quadra>
  Olha, <realçado> Daisy </realçado> : quando eu morrer tu hás-de
  dizer aos meus amigos aí de <realçado> Londres </realçado>,
  embora não o sintas, que tu escondes
  a grande dor da minha morte. Irás de
</quadra>
```

Anotação orientada à estrutura

```
<SEC>Isto é uma secção de nível 1.
<SEC>Isto é uma secção de nível 2.</SEC> </SEC>
<P>Isto é um parágrafo do nível de topo.</P>
<LISTA>
  <ITEM>Isto é um item de uma lista de nível 1.
  <LISTA><ITEM>Isto é um item de uma lista de
  nível 2.</ITEM>
</LISTA></ITEM></LISTA>
```

```
<SEC1>Isto é uma secção de nível 1.</SEC1>
<SEC2>Isto é uma secção de nível 2.</SEC2>
<P0>Isto é um parágrafo do nível de topo.</P0>
<LISTA1><ITEM>Isto é um item de uma lista de nível 1.</ITEM></LISTA1>
<LISTA2><ITEM>Isto é um item de uma lista de nível 2.</ITEM></LISTA2>
```

Anotação orientada ao conteúdo

```
...
<quadra>
  <verso>Olha, <nome>Daisy</nome>: quando eu morrer
    tu hás-de</verso>
  <verso>dizer aos meus amigos aí de
    <lugar>Londres</lugar>,</verso>
  <verso>embora não o sintas, que tu escondes</verso>
  <verso>a grande dor da minha morte. Irás de</verso>
</quadra>
...
```

Anotação Equilibrada

- Exemplo: DocBook
  - Formato: EMPH, TABLE,...
  - Estrutura: SECT1, SECT2, SECT3, ...
    - Conteúdo: NAME, AUTHOR, PUBDATE, COMMAND, ...

Documentos XML bem formados

Um documento XML

- Conteúdo = Dados + Anotações
- Dados = blocos de texto
- Anotações:
  - marcas de início de elementos
  - marcas de fim de elementos
  - marcas de elementos vazios
  - referências a entidades
  - comentários
  - limitadores de secções especiais de texto
  - declarações de tipo de documento
  - instruções de processamento

O exemplo tradicional

```
<?xml version="1.0" encoding="UTF-8"?>
   <doc>
         Hello World!!!
   </doc>
```

A declaração XML

- Anotação especial que deve iniciar todos os documentos XML

```
<?xml
    version="1.0"
    standalone="yes"
    encodi
```

version - obrigatório, valores possíveis: 1.0

standalone - opcional, valores possíveis: yes, no;

encoding - opcional, para o português o valor deverá ser: ISO-8859-1

Comentários

- Podem aparecer em qualquer ponto dum documento XML.
- Começam pela marca: <!--
- e terminam com a marca: -->.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!--Isto é um comentário no início-->
<doc>
    Olá Mundo!!!
</doc>
```

Comentários (2)

- Existem algumas restrições à utilização de comentários:
  - Não podem aparecer antes da declaração.
  - Não podem aparecer dentro duma anotação.
  - Não se pode utilizar a sequência de carácteres "--" dentro dum comentário.

Comentários (3)

- Podem ainda ser utilizados para remover temporariamente partes do documento, desde que essas partes não contenham comentários.

```
<RECEITAS>
 <TITULO> O Meu Livro de Receitas </TITULO>
 <RECEITA ORIGEM="Portugal">
  <TITULO> Bolo </TITULO>
 <!--
  <INGREDIENTE> 500g de farinha </INGREDIENTE>
 -->
  <INGREDIENTE> 200g de açucar </INGREDIENTE>
  <INGREDIENTE> 300g de manteiga
</INGREDIENTE>
 </RECEITA>
</RECEITAS>
```

Instruções de Processamento

- As instruções de processamento são uma reminiscência da anotação procedimental
- Uma instrução de processamento não faz parte do conteúdo do documento.
- É uma indicação directa de que algo deve ser executado naquele ponto.
- Uma instrução de processamento começa por: <?id-processador
- e termina por: ?>
- Exemplo: a declaração XML

Instruções de Processamento 2

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<agenda>
<?html action="hr"?>
    <entrada id="e1" tipo="pessoa">
        <nome>José Carlos Ramalho</nome>
        <email>jcr@di.uminho.pt</email>
        <telefone>253 604479</telefone>
    </entrada>
<?html action="hr"?>
...
</agenda>
```

Elementos

- Blocos lógicos em que um documento pode ser decomposto
- Exemplo:
  - **Vais ver o espectáculo a <lugar>Braga</lugar>?**
- Uma anotação de início começa por **<** e termina por **>**,
- e uma anotação de fim começa por **</** e termina por **>**.
- Uma anotação contém o nome do elemento que inicia ou que termina, respectivamente.

Caracteres reservados

- No conteúdo dum elemento, nunca deverão ser usados os caracteres '**<**' e '**>**' pois são os caracteres que limitam as anotações.
- Em lugar deles devem-se usar, respectivamente, as entidades do tipo carácter '**<**' e '**>**'.
- Qualquer processador ou editor de XML fará a substituição automática daquelas entidades pelos caracteres correspondentes.

Caracteres reservados 2

Se, no documento, estivesse o seguinte texto:

A anotação <nome> é usada para anotar nomes.

Um editor mostraria o mesmo texto da seguinte maneira:

A anotação <nome> é usada para anotar nomes.

Tipos de Conteúdo

- elemento com conteúdo textual
  - <lugar>Braga</lugar>
  - <INGREDIENTE>Meia dúzia de ovos</INGREDIENTE>
  - <data>(1922)</data>
- elemento com conteúdo misto
  - **<verso>**Olha, <nome>Daisy</nome>: quando **...</verso> **Vais ver o espectáculo a <lugar>Braga</lugar**>? **

Tipos de Conteúdo (2)

- elementos com conteúdo vazio: são normalmente utilizados pelo seu significado posicional - referências, pontos de inserção de imagens, ...
  - **Como será discutido num capítulo mais à frente (<ref ident="cap5"/>) ...**
  - São representados por uma única anotação que é iniciada por '<' e termina em '/>', que é a forma abreviada de escrever “<elem-ident></elem-ident>”.

Atributos

- Um elemento pode ter um ou mais atributos que, por sua vez, podem ser opcionais ou obrigatórios.
- Visam qualificar o elemento a que estão associados.
- Não há limite para o número de atributos que podem estar associados a um elemento.
- Aparecem sempre na anotação que marca o início dum elemento, uma vez que vão qualificar o conteúdo que se segue.

Atributos (2)

- Um atributo é definido por um par constituído por um nome e um valor:
  - o nome e o valor devem estar separados pelo sinal ‘=‘ e
  - o valor deverá estar colocado dentro de aspas simples ou duplas.
  - Exemplo:

    <ref destino=“exemplo5”/>

    <imagem path=“figs/img3.gif”/>

Elemento versus Atributo

- Não existe uma fronteira entre os dois e muitas vezes a escolha não é simples.

Informação nos elementos

```
<agenda>
  <entrada id="e1" tipo="pessoa">
    <nome>José Carlos Ramalho</nome>
    <email>jcr@di.uminho.pt</email>
    <telefone>253 604479</telefone>
  </entrada>
  ...
</age
```

Informação nos atributos

```
<agenda>
  <entrada id="e1" tipo="pessoa" nome="José Carlos Ramalho"
           email="jcr@di.uminho.pt" telefone="253 604479"/>
  ...
</agenda>
```

Atributos reservados

- xml:lang
  - Pode ser
    indica qua
    elemento

```
<para xml:lang="en">Hello</para>
<para xml:lang="pt">Olá</para>
<para xml:lang="fr">Bonjour</para>
```

- xml:space
  - Serve para indicar se o espaço branco no conteúdo do elemento em causa é ou não relevante: **preserve** ou **default**.

Secções Marcadas

- Úteis para incluir exemplos de XML

```
<![ CDATA [
The <p> tag is used for paragraphs
]]>
```

  - Ou para texto com muitos caracteres reservados:

```
Prima a tecla <<<ENTER>>>.

Prima a tecla &lt;&lt;&lt;ENTER&gt;&gt;&gt;.
<![CDATA[Prima a tecla <<<ENTER>>>.]]>
```

Regras de bem-formação

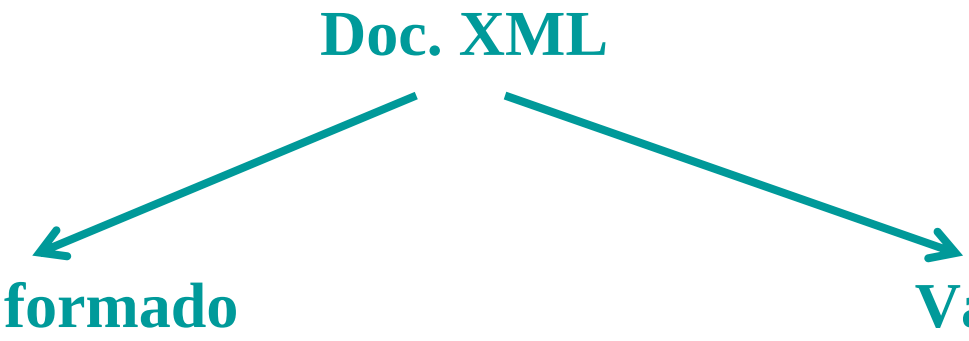
- Um documento XML deve ter sempre uma declaração XML no início
- O documento deve incluir um ou mais elementos
- Todos os elementos têm anotações de início e fecho (excepto os vazios)
- Os elementos deverão estar aninhados correctamente
- Os valores de atributos têm de estar dentro de aspas

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<sumarios>
  <disciplina> Processamento Estruturado de
               Documentos</disciplina>
  <professor>
     <nome>José Carlos Ramalho</nome>
     <email>jcr@di.uminho.pt</email>
     <url>http://www.di.uminho.pt/~jcr</url>
  </professor>
  <aula tipo="T">
     <data>2000.10.02</data>
     <sumario>
        <p>
          Anotação de Documentos: um pouco de
          história.</p>
        <p>
          Linguagens de Anotação como meta-linguagens:
          o SGML e o XML.</p>
        <p>
          Anotação Descritiva. Ciclo de vida dos
          documentos estruturados.</p>
     </sumario>

  </aula>
  ...
</sumarios>
```

Documento bem-formado

XML (conceitos)



- não há cruzamento de tags

  <A>olá <B>estás </A> ...
- pode-se inferir um DTD ou um Schema
- torna o pós-processamento

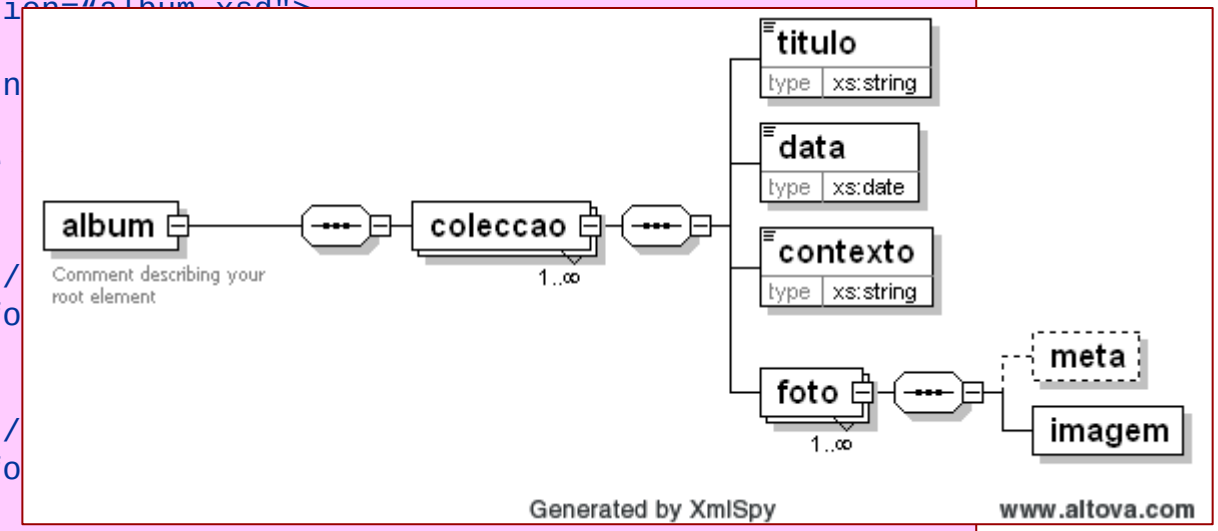
mais específico.

- pertence a uma classe (DTD)

Documentos XML válidos

Um documento XML é válido quando a sua estrutura/sintaxe é verificada e está de acordo com um **DTD** ou um **Schema**.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<album xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="album.xsd">
<coleccao>
  <titulo>Orientação Noturn
  <data>2004-06-03</data>
  <contexto>Actividades de
Minho</contexto>
  <foto>
    <meta path="meta1.xml"/
    <imagem path="foto1" fo
  </foto>
  <foto>
    <meta path="meta2.xml"/
    <imagem path="foto2" fo
  </foto>
  ...
```



Especificação de DTDs

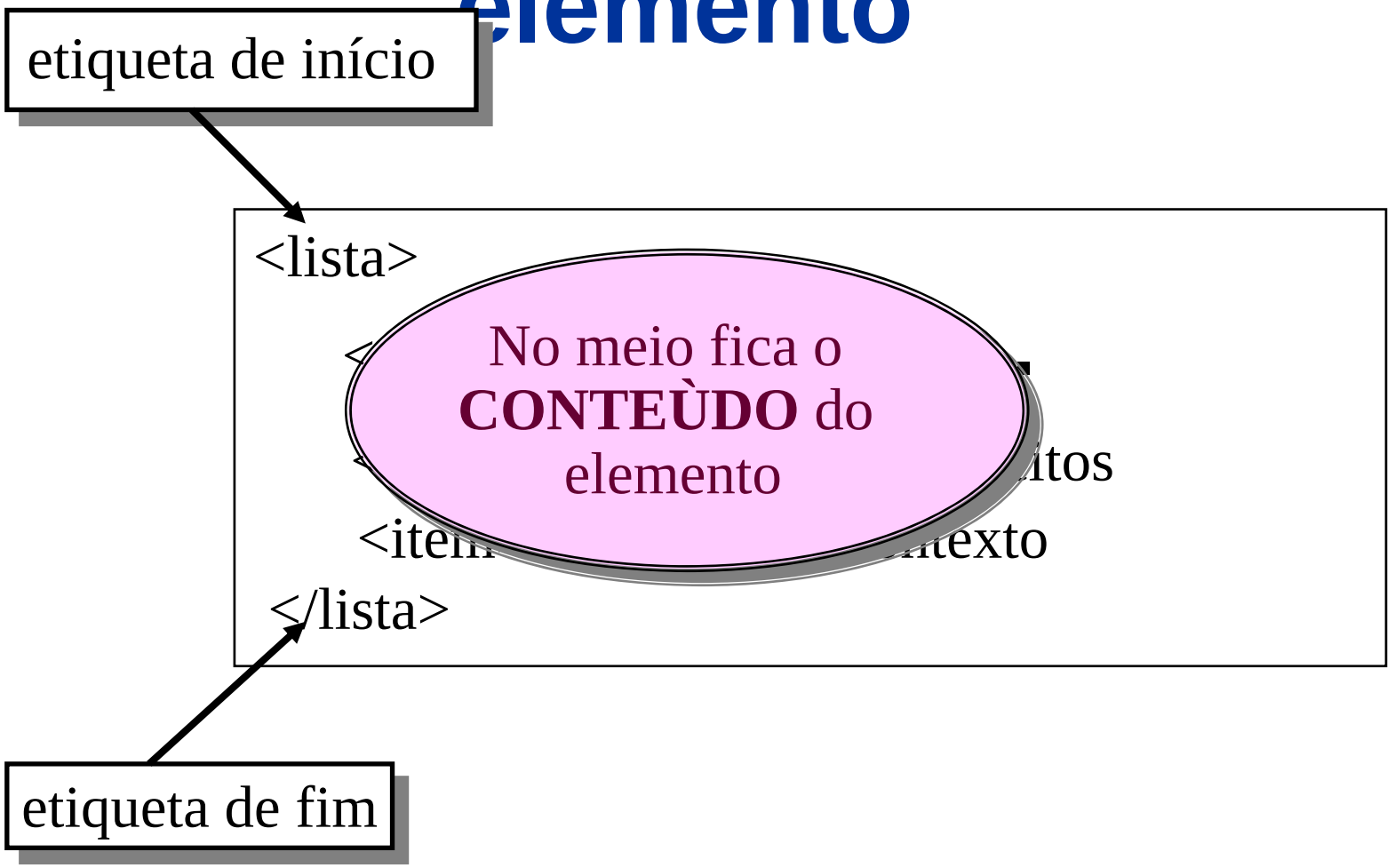
- DOCTYPE
- ELEMENT
- ATTLIST
- ENTITY

XML: exemplo

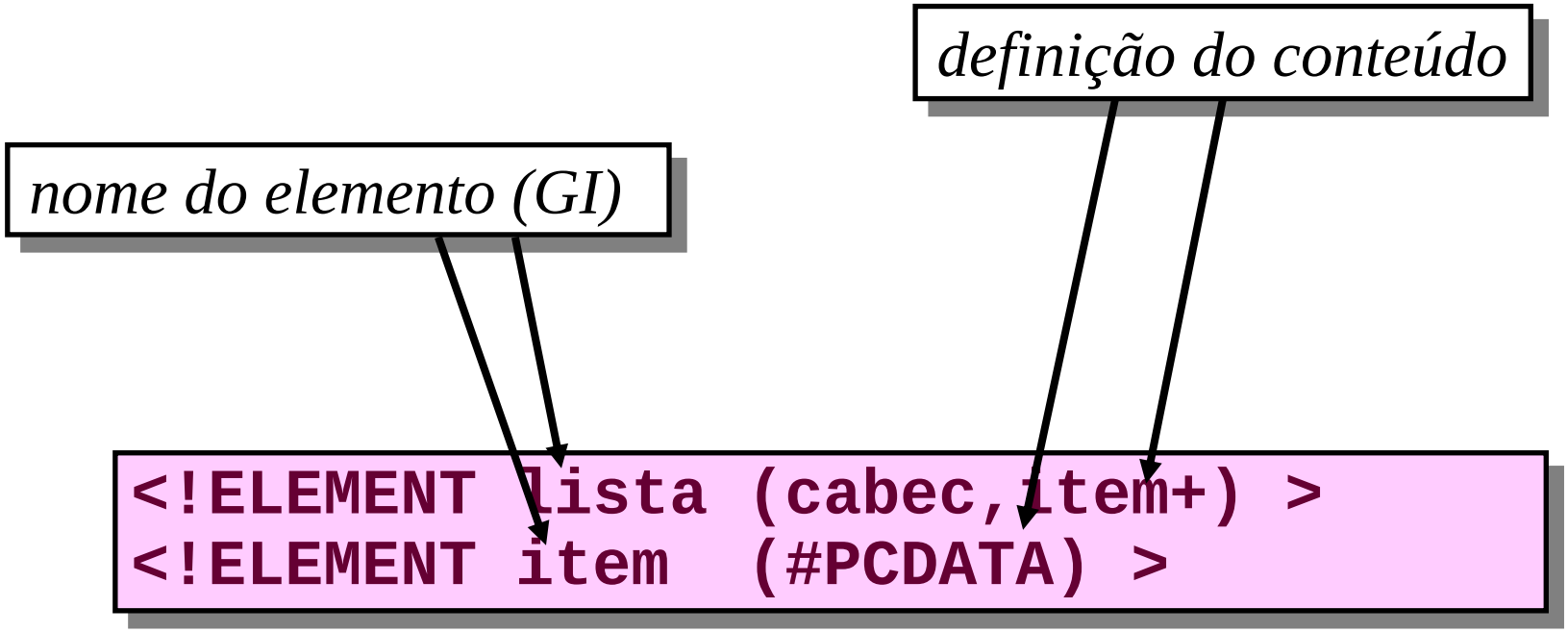
```
<lista>
  <cabec>Os delimitadores podem ser: </cabec>
  <item> explícitos </item>
  <item> inferidos do contexto </item>
</lista>
```

*O elemento do tipo **LISTA** é formado por um elemento **CABEC**, seguido por dois elementos do tipo **ITEM**.*

Ocorrência de um elemento



Definição de um elemento



Definição do CONTEÚDO

- outros elementos especificados
- ANY (qq elemento especificado)
- EMPTY (nada, vazio)
- #PCDATA (texto)
- uma mistura de elementos com #PCDATA

Expressão de Conteúdo: sintaxe

- sequência
  - a,b — a *seguido de* b
  - a|b — a *ou* b *mas não ambos*
- ocorrência
  - a — *um e apenas um*
  - a? — *opcionalmente um (0 ou 1)*
  - a* — *zero ou mais*
  - a+ — *um ou mais*

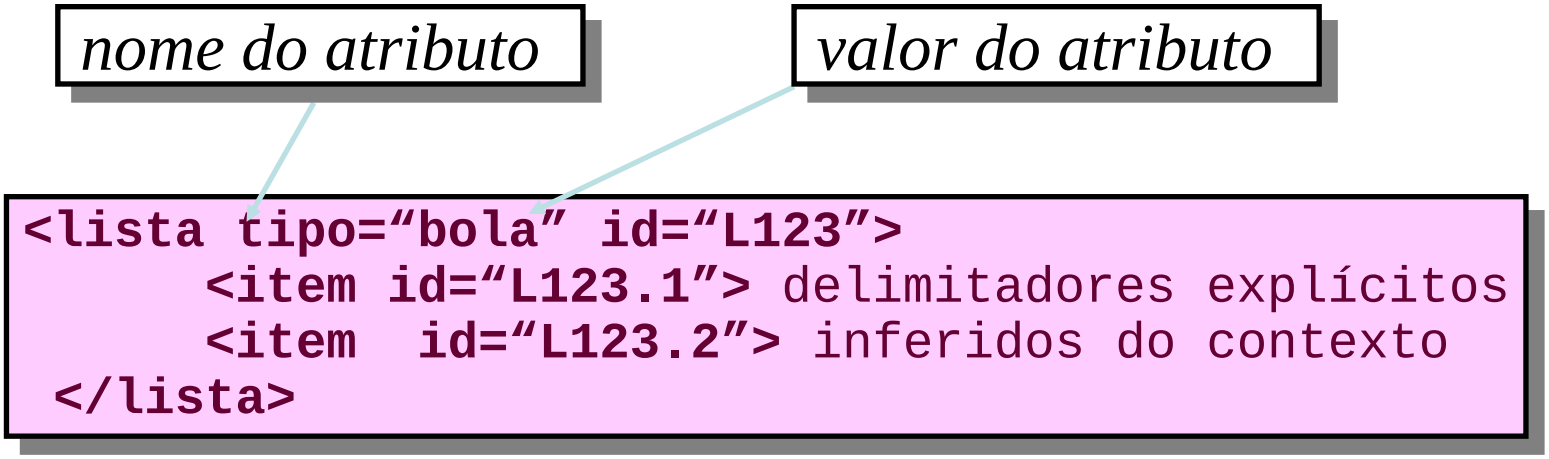
Exemplo: o poema

Poema: título, autor, corpo, data
Corpo: quadra, quadra, terno, terno.
Quadra: verso, verso, verso, verso
Terno: verso, verso, verso
Verso: (texto | nome)+

```
<!ELEMENT poema (titulo,autor,corpo,data) >
<!ELEMENT corpo (quadra,quadra,terno,terno) >
<!ELEMENT quadra (verso,verso,verso,verso) >
<!ELEMENT terno (verso,verso,verso) >
<!ELEMENT verso (#PCDATA |nome)* >
```

Um elemento pode ter atributos

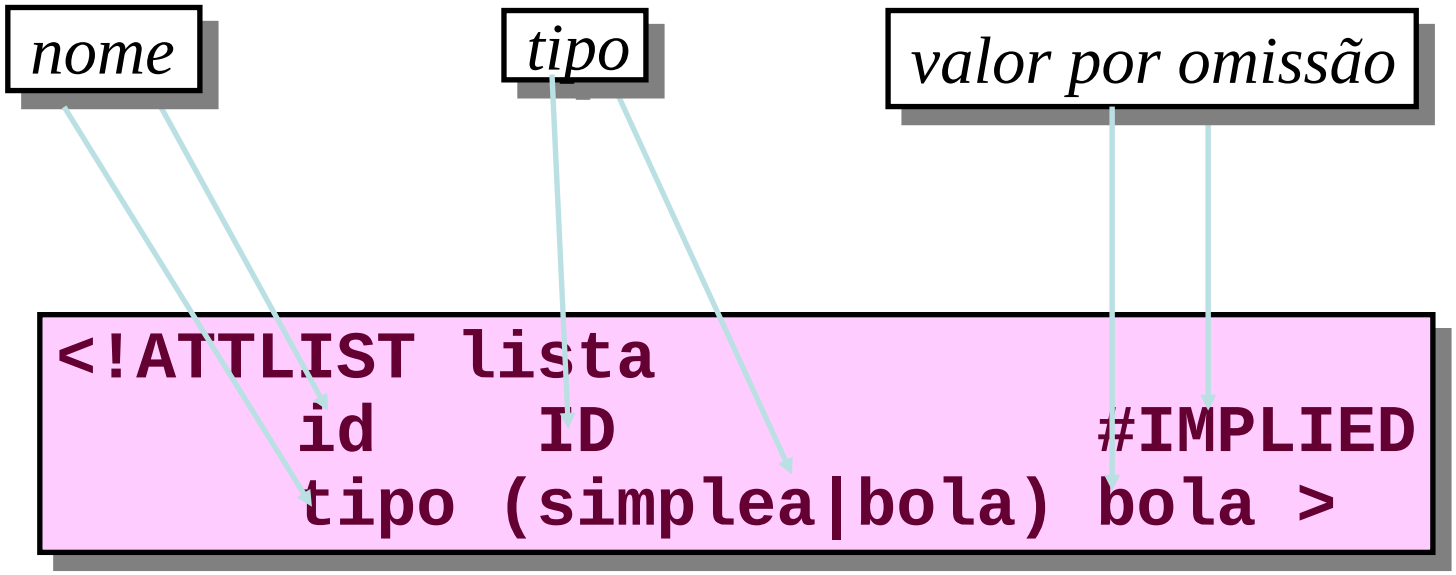
- para conter informação para além do tipo e do contexto
- para identificação de ocorrências específicas de elementos
- para fazer algumas validações (poucas)



Ex: anotação morfo-sintáctica

<quadra>
<verso><**verbo tempo="imperativo" pessoa="2s**">Olha</**verbo**>,<nome> Daisy</nome>: quando eu morrer tu hás-de</verso>
<verso><**verbo tempo="infinitivo**">dizer</**verbo**> aos meus amigos aí de <nome>Londres</nome>,</verso>
<verso>embora não o sintas, que tu escondes</verso>
<verso>a grande dor da minha morte. Irás de</verso>
</quadra>

Definição de um atributo



os identificadores de nomes e tipos devem ser únicos dentro dum elemento

Tipos de atributo

- **ID** *um identificador único dentro o documento actual*
- **IDREF** *referência a um identificador definido algures no documento corrente*
- **CDATA** *texto*
- **NAME, NUMBER, NMTOKEN**
- **ENTITY** *o nome duma entidade definida no documento corrente*
- *uma enumeração/lista de valores específicos (não pode haver repetições na lista)*

Valores possíveis

- #REQUIRED (obrigatório)
- #FIXED (constante)
- #IMPLIED (opcional)
- *valor explícito*