Introdução ao XML

Roteiro da Aula

- Dados Semi-estruturados
- O que é XML
- XML x HTML
- Terminologia XML
- Namespaces

Dados estruturados ou não...

- Dados estruturados
  - Estrutura é conhecida a priori
    Ex.: Dados de um SGBD relacional têm um esquema relacional associado
- Dados não estruturados
  - Não há nenhuma estrutura prévia
    Ex.: imagem, video, áudio, etc.

Dados Semi-estruturados

- Dados irregulares
  - Livros podem ser descritos por uma estrutura de partes e capítulos ou podem ser descritos somente por capítulos.
  - A descrição de uma disciplina pode variar em termos de atributos de um departamento para outro:
    - faltam atributos ou apresentam atributos a mais
- Dados incompletos
  - Nem todo endereço tem caixa postal
  - Nem todo livro tem apêndice ou prefácio
- Não necessariamente está de acordo com um esquema
  - Sua estrutura não é previamente conhecida, não existe à parte
  - São auto-descritivos, i.e., embute a própria estrutura.

Dados Semi-estruturados
Como se auto-descrevem...

– pares atributo-valor

{name: “John Smith”, tel: 3456, age: 32}

– valor de atributo pode também conter estrutura

{name: {first:“John”, last: “Smith”},

tel: 3456, age: 32}

– rótulos de atributo não necessariamente únicos

{name: “John Smith”, tel: 3456, tel: 7891}

Dados Semi-estruturados

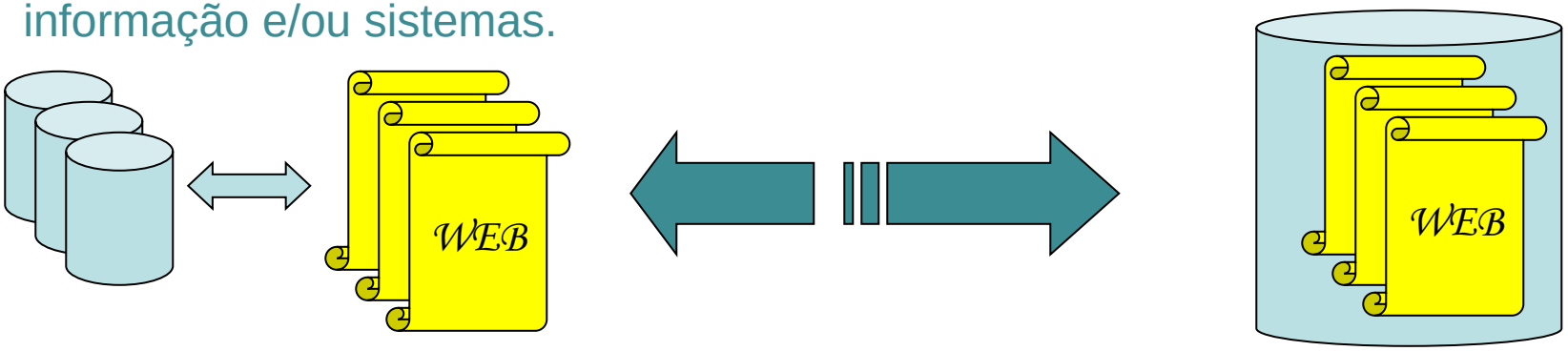
- Situações típicas
  - Qdo os dados não podem ser restritos a um esquema
    - Difícil definir uma estrutura... Ex: contratos
  - Qdo não há compromisso com o conteúdo
    - Pode-se ter muitos dados faltando... Ex. Leis
  - Qdo as fontes de dados são heterogêneas e é preciso integrar dados...
    - Descrições equivalentes mas distintas...

Web: grande fonte de dados semi-estruturados

- Páginas web contém informação valiosa
  - Documentos de conteúdo importante
  - Dados armazenados em BD's disponibilizados na web
- Novas aplicações surgem com outro objetivo
  - Intercambiar e/ou extrair informação da web
  - Monitoração do acesso/navegação do usuário

Antes, a web era vista como uma forma de disponibilizar informação e/ou sistemas.

Hoje, a Web é vista como um grande banco de dados.



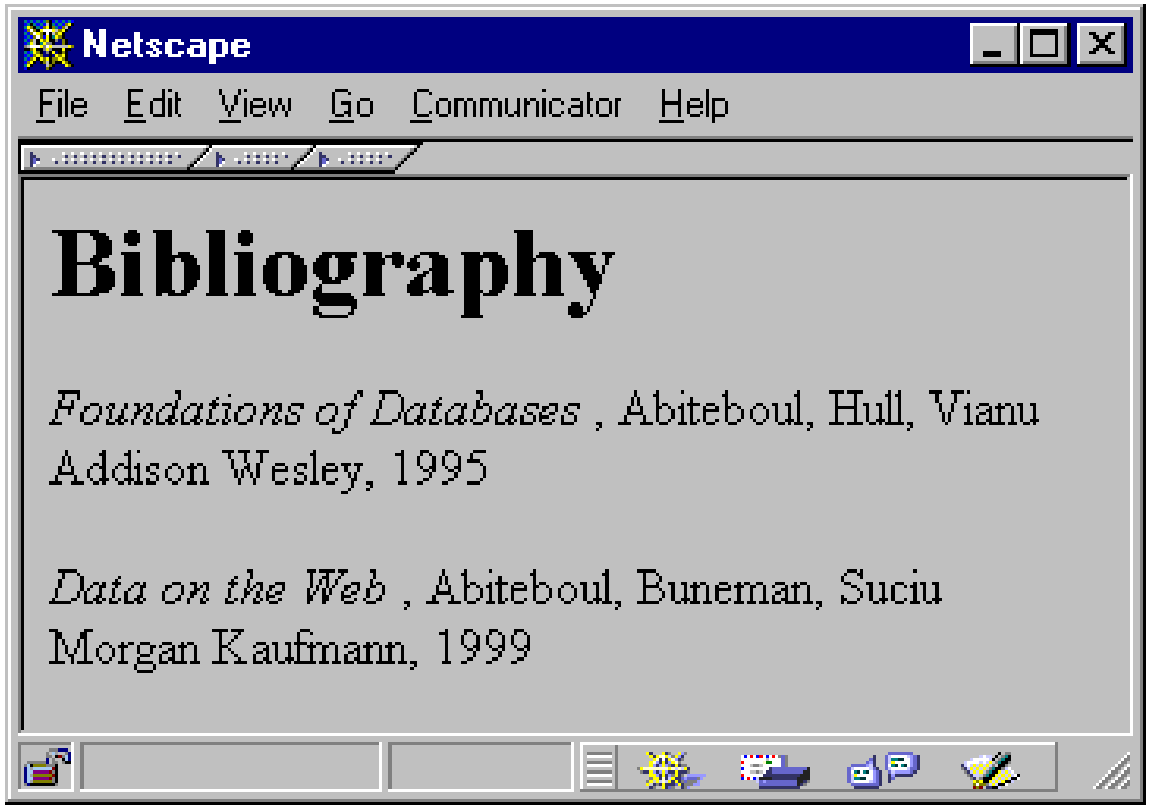
Descrever os dados da Web

- Tratar dados semi-estruturados
- Separar o conteúdo:
  - Independência de armazenamento
    - Permite a visualização de dados provenientes de fontes heterogêneas
  - Independência de apresentação
    - Permite que as aplicações apresentem/tratem os dados como lhes é conveniente

O que é XML?

- eXtensible Markup Language
- Padrão para marcação de dados na Web, com foco na descrição do conteúdo – W3C (ww.w3c.org)
- HTML – descreve o **formato** do documento
  - HTML tem um conjunto fixo de tags e não descreve conteúdo
- XML – descreve o **conteúdo** do documento
  - Usuário define suas próprias tags para criar uma estrutura
  - Um documento XML não tem nenhuma instrução para apresentação

De HTML para XML...



HTML descreve a **apresentação!**

Fonte HTML

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<HTML>
  <HEAD><TITLE>A bibliography on Databases</TITLE>
  <META content="text/html; charset=windows-1252" http-equiv=Content-Type>
  <META content="MSHTML 5.00.2314.1000" name=GENERATOR>
  </HEAD>
  <BODY>
      <h1> Bibliography </h1>
      <p> <i> Foundations of Databases </i> Abiteboul, Hull, Vianu <br>
          Addison Wesley, 1995
      <p> <i> Data on the Web </i> Abiteoul, Buneman, Suciu <br>
          Morgan Kaufmann, 1999

  </BODY>
</HTML>
```

HTML: Conjunto pré-definido de elementos (tags) para especificação das dimensões de estrutura e apresentação de um documento

Fonte XML

```
<bibliography>
      <book>    <title> Foundations... </title>
                <author> Abiteboul </author>
                <author> Hull </author>
                <author> Vianu </author>
                <publisher> Addison Wesley </publisher>
                <year> 1995 </year>
      </book>
      ...
</bibliography>
```

XML: Elementos (tags) definidos pelo usuário da linguagem e servindo para descrever o conteúdo e a estrutura.

XML descreve o **conteúdo**!!!

Dimensões de informações em um documento

- Documentos apresentam pelo menos duas dimensões de informações:
  - o conteúdo propriamente dito
  - a estrutura organizacional

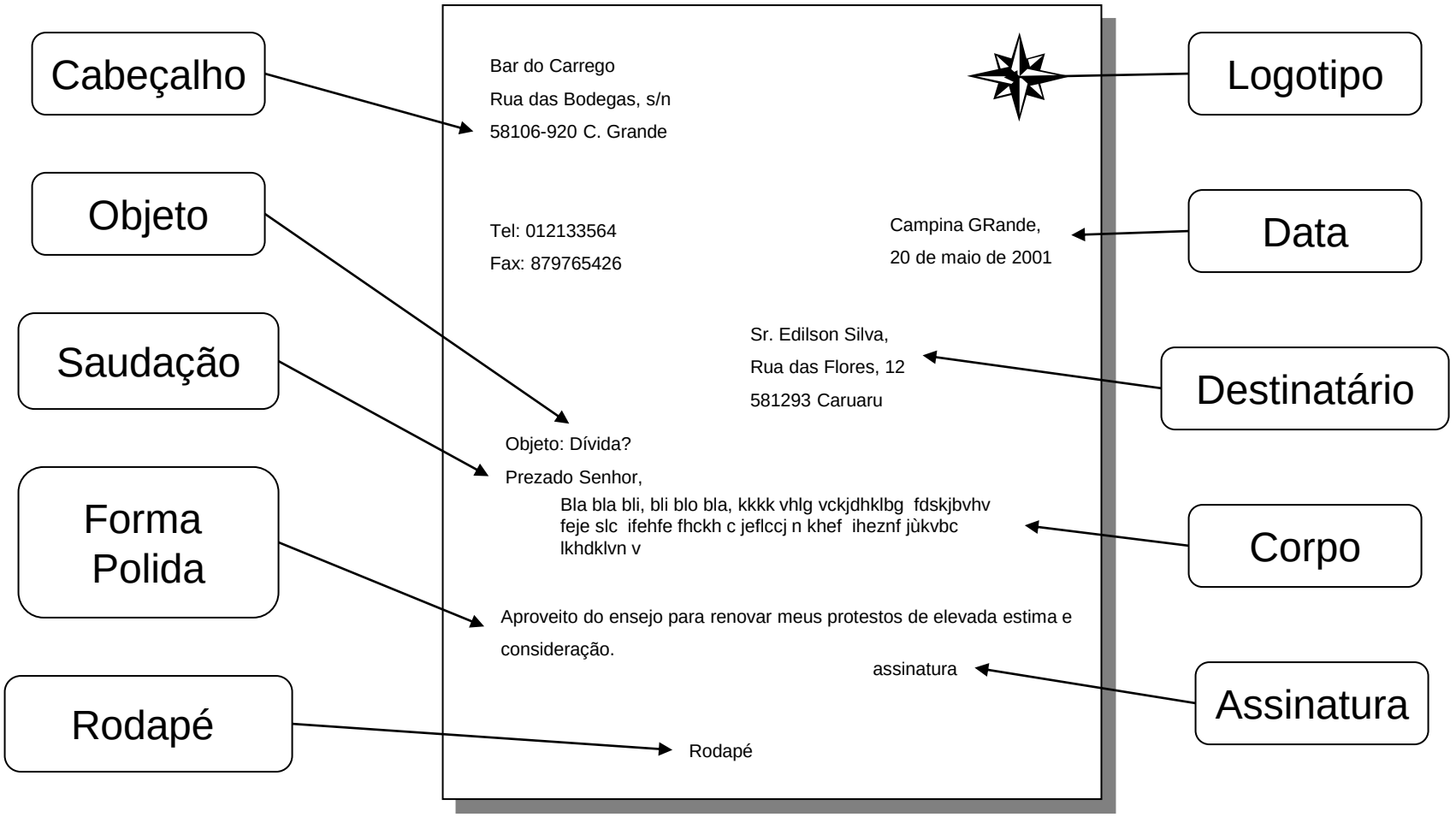
XML: dimensões e processamento ...

- XML
  - Dimensões de estrutura e conteúdo
  - Documentos bem formados!
- Outras dimensões de um documento XML
  - Apresentação: CSS, XSL
  - Mais estrutura e semântica: DTDs e XML Schemas
  - Metadados e mais semântica: RDF
  - Estrutura de hipertexto: XLink e XPointer
- Processamento de documentos XML
  - Parsers, APIs, DOM...
  - Aplicações em geral

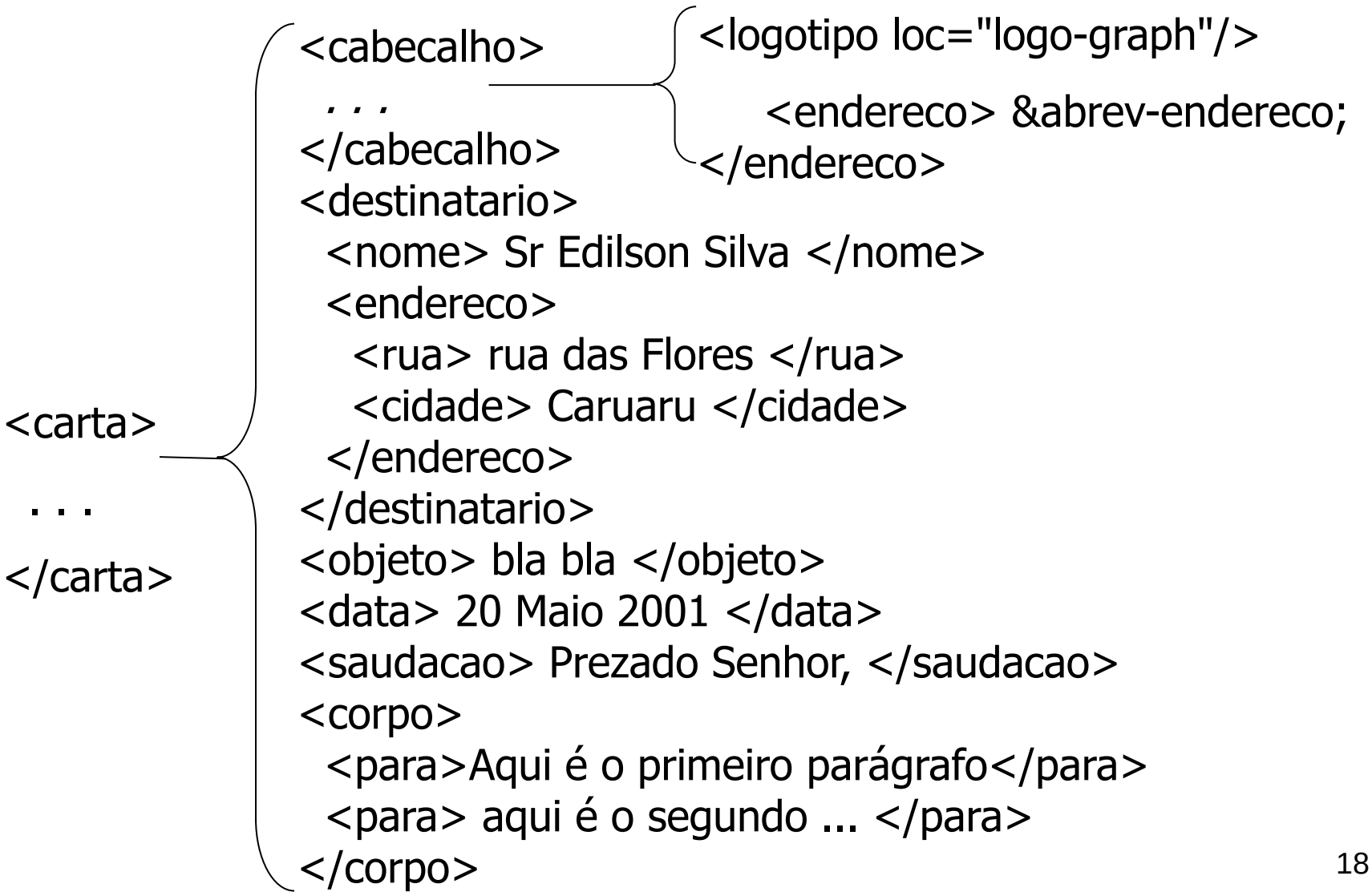
E a apresentação?

- Uma representação em XML não tem diretamente nenhuma informação de apresentação.
- As numerosas propriedades gráficas ou tipográficas estão ausentes da fonte XML.
- Estas propriedades serão definidas por intermédio de um informações suplementares, em uma folha de estilo associada ao documento XML
- Uma folha de estilo é um *conjunto de regras* para especificar a *realização concreta* de um documento sobre uma *mídia* particular.

Exemplo de um documento



Representação XML

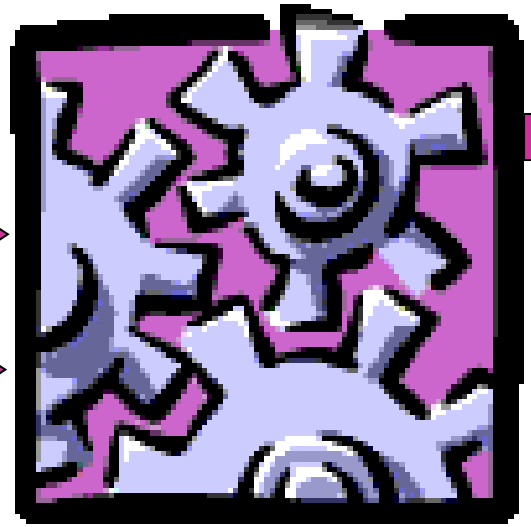


Princípio de funcionamento das folhas de estilos

```
<carta>
  <cabecalho>
    . . .
  </cabecalho>

  <corpo>
    . . .
  </corpo>
</carta>
```

If `carta` then . . .
If `cabecalho` then ...
If `corpo` then
...
If `para` then
`Times new roman,`
`size 12,`
`skip first line`
If ... then ...



WindStar 2000

Les rosières en buget

AB562 Saint Pétaouchnoque

Tel: 012133564

Fax: 879765426

Saint Pétaouchnoque,

Le 30 nivose 2004

Editions Duschmol,

12 rue Schmurz

YT123 Rapis

Objeto: Dívida

Prezado Senhot,

Bla bla bli, bli blo bla, kkkk vhlg vckjdhklbg fdskjbvhv feje slc ifehfe fhckh c jeflccj n khef iheznf jùkvbc lkhdklvn v

.

sssinatura

Rodapé

Por quê XML?

Extensibilidade e estrutura

- Em XML, um autor ou uma comunidade de autores inventam livremente as tags que lhes pareçam úteis para marcar os componentes de um documento.
- Exemplo: diversas formas de representar uma data
  - <date> 5 janeiro 2000 </date>
  - <date>
    <ano> 2000 </ano>
    <mes> 01 </mes>
    <dia> 05 </dia>
    </date>
  - <date format='ISO-8601'> 2000-01-05 </date>
- Grande liberdade de escolha das estruturas de dados facilita ***a troca de dados***

Interoperabilidade

- Todos os dados podem ser vistos como documentos XML e não mais como arquivos no formato X ou Y.
- Consequências:
  - Um servidor de documentos XML é suscetível de responder a um conjunto de necessidades de uma organização.
  - Um simples editor de textos pode tratar o conjunto de dados de uma organização.
  - A interoperabilidade dos utilitários está assegurada.

Modularidade e reutilização

- Cada usuário é livre para definir suas próprias estruturas de documento
- Ele pode também estar conforme as estruturas tipadas, chamadas DTD
- Cada comunidade pode propor as estruturas normalizadas
- A validação a uma DTD permite a automatização no tratamento dos dados e assegura uma possibilidade de controle de integridade

Acesso à fontes de informação heterogêneas

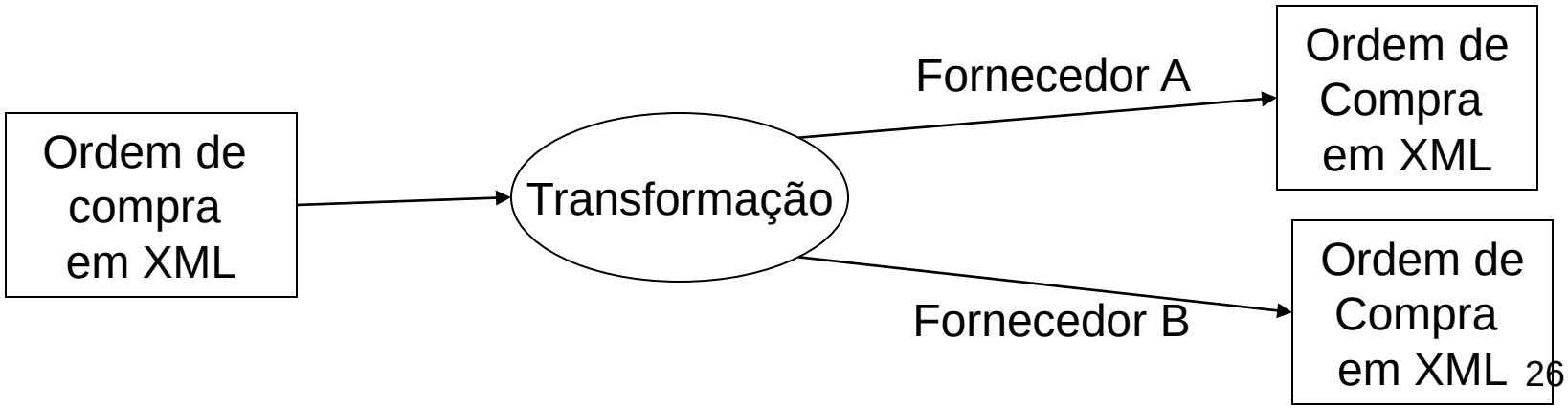
- A consulta e troca de dados entre as base de dados heterogêneas é complexa
  - XML contribui para minimizar este problema: formato de troca normalizado, genérico, independente de plataforma

Acesso à fontes de informação heterogêneas

- A indexação e consulta de bases de documentos pode se beneficiar de informações estruturais e textuais.
  - pesquisa por palavras-chaves: Jorge+Amado retorna todos os documentos contendo as palavras Jorge e Amado
  - pesquisa estrutural: pesquisa os documentos cujo autor é Jorge Amado (ie os documentos contendo um elemento `autor`, ou `escrito-por` contendo Jorge e Amado)

XML no Mercado

- *XML está se tornando uma plataforma padrão para os processos entre empresas dos quais depende o comércio eletrônico B2B. – W. Lewis*
  - B2B e-commerce: empresas que centralizam múltiplos vendedores e compradores, compatibilizando ordens de compra e venda entre eles.



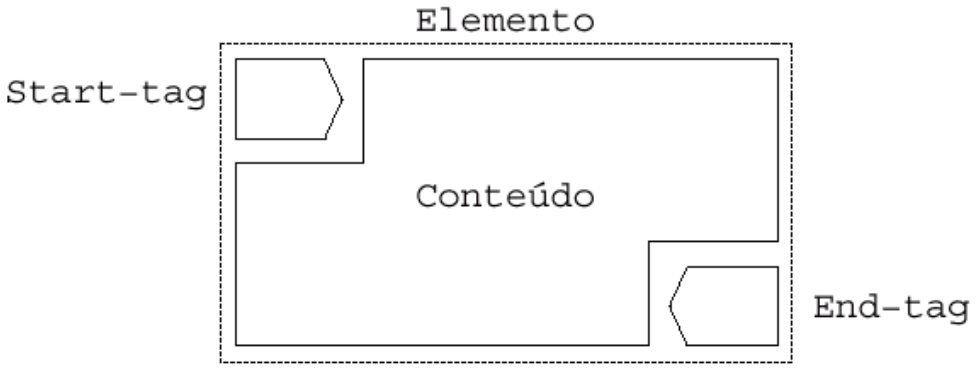
XML - Apenas uma linguagem de marcação?

- A linguagem XML tem associada uma série de iniciativas:
  - XSL
  - SAX, DOM
  - DTD, XML Schema
  - XLink and XPointer
  - XPath, XQuery
  - RDF, OWL
  - Serviços Web (WSDL, etc.)
  - etc.

Sintaxe XML

Marcação XML

- Instruções de marcação XML são denominadas *tags*
  - **Não** especificam um estilo de apresentação particular
  - **Identificam** a natureza de um componente de texto
- As *tags* XML delimitam um objeto identificável no fluxo de dados



- Um elemento XML é formato por uma *start-tag* (marca inicial) o conteúdo propriamente dito e uma *end-tag* (marca final) – *OBRIGATORIAMENTE*

Marcação XML

- A *start- tag* e a *end- tag* encontram- se distribuídas no fluxo de dados
- Objetivo de delimitar objetos identificáveis
- Exemplo:

<pergunta>
Vai viajar para <cidade> Porto Alegre </ cidade> segunda?
</pergunta>

Porto Alegre É uma cidade, assim a *tag* usada para demarcar a informação recebe o nome de cidade

Tipos de marcas

– Composta

```
<from>
   <name>Carina Dorneles</name>
   <email>dorneles@inf.ufrgs.br</email>
</from>
```

Marca composta de outras marcas

– Texto

```
   <name>Carina Dorneles</name>
```

Marca composta de conteúdo texto

– Mista

```
<from>Carina Dorneles
   <email>dorneles@inf.ufrgs.br</email>
</from>
```

Marca composta de outras marcas + conteúdo texto

– Vazia

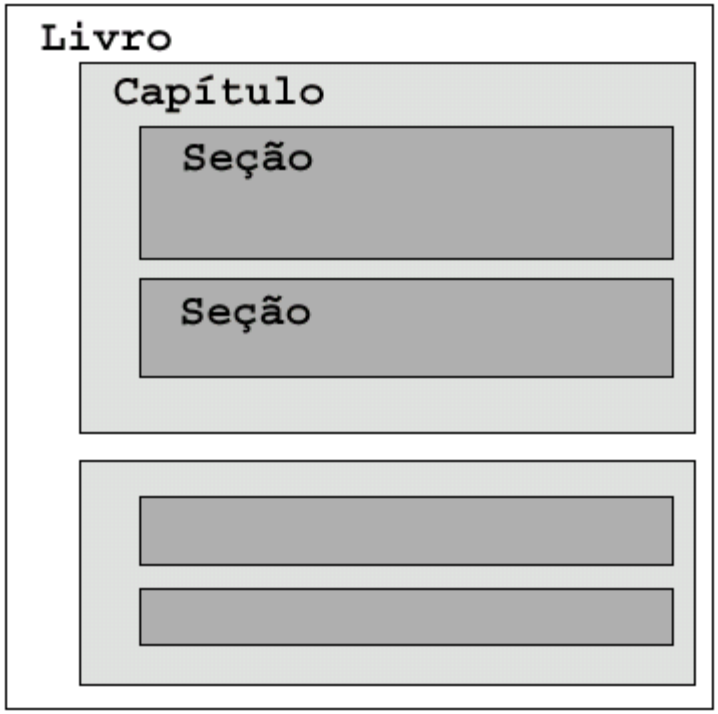
```
<profissao><engenheiro/></profissao>
```

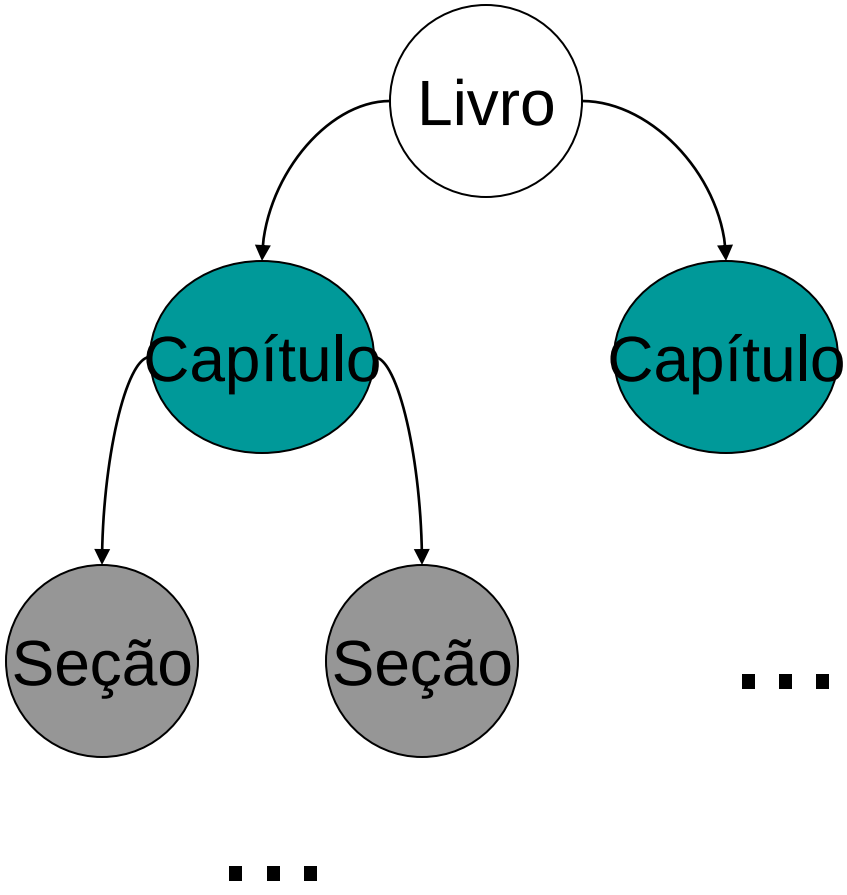
Marca sem conteúdo

Estrutura de um documento XML

- Elementos podem conter outros elementos
  - Aninhamento de *tags*
- O documento completo encontra-se delimitado por *um único elemento* denominado **"elemento documento"** ou **"elemento raiz"**
- Pode ser visualizada de diferentes formas
  - Forma de árvore, caixas dentro de caixas, etc.

Visualização da estrutura





Regras de estruturação

- Cada elemento possui um único pai
- Cada elemento possui um número arbitrário de irmãos e filhos
  - Um elemento sem filhos é denominado *folha*
- Exceção: o **elemento documento/elemento raiz** não possui pai e não possui irmãos

Regras de estruturação

- Todas as *tags* devem ser fechadas:

```
<p> Parágrafo em HTML
<p> um possível parágrafo em XML </p>
```

- As tags XML são case sensitive

```
<Mensagem> Isto está incorreto </mensagem>
<MENSAGEM> Isto é correto </MENSAGEM>
<mensagem> Isto é correto </mensagem>
```

Regras de estruturação

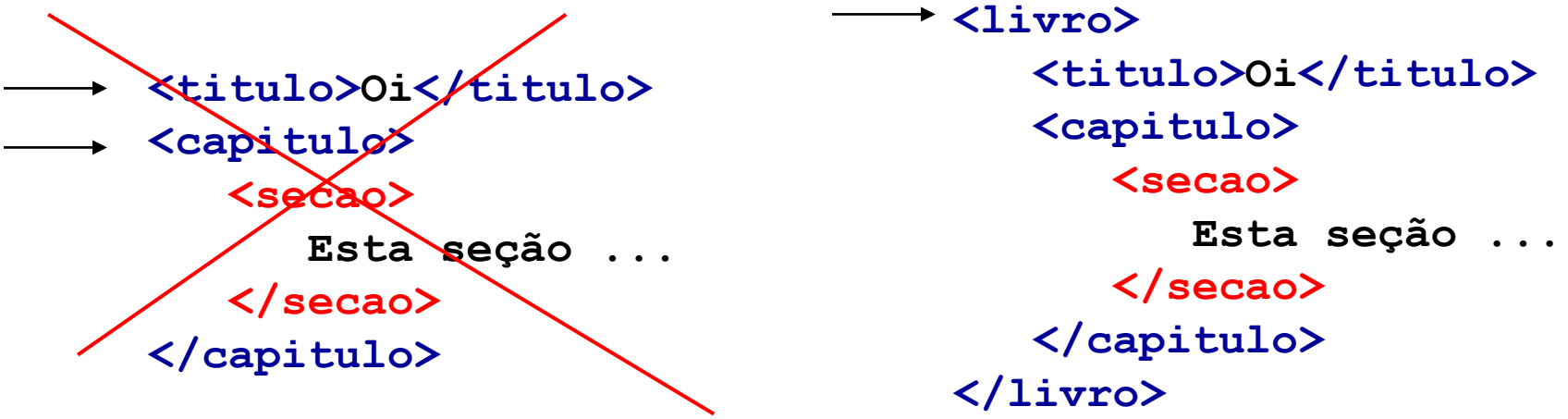
- As *tags* devem estar bem aninhadas

```
<b>
  <i>Isto é permitido em HTML, mas não em XML</b>
</i>
```

```
<b>
  <i> Isto é permitido em XML </i>
</b>
```

Regras de estruturação

- TODO documento XML deve possuir uma raiz



- Valores de atributos devem estar entre aspas

```
<mensagem data="20.10.2001">
  Esta é uma mensagem enviada em 20 de outubro de
  2001
</mensagem>
```

Estruturas hierárquicas

- É possível definir regras que expressam como os elementos podem ser combinados (estrutura hierárquica)
  - DTD *(Document Type Definitio*n)
  - XML Schema

- Estruturas hierárquicas recursivas
  - Um elemento pode conter direta ou indiretamente instâncias de si mesmo

    Um elemento pode conter direta ou indiretamente instâncias de mesmo tipo (uma seção pode conter outras seções

Estruturas hierárquicas

- Estruturas hierárquicas recursivas

```
<list>
  <item>...</item>
  <item>...</item>
  <item>
    <list>
      <item>...</item>
      <item>...</item>
    </list>
  </item>
</list>
```

Atributos

- Um elemento pode conter um **número arbitrário** de atributos
  - Cada atributo é um par (nome, valor), separados por “=“
  - Os valores dos atributos são cadeias de caracteres e devem ser delimitadas por aspas
  - Um atributo possui um tipo quando é utilizada uma DTD
  - Pode-se associar um valor *default* a um atributo
    - O que não se pode fazer com um elemento
    - Isso é feito através de uma DTD

Atributos

- Valores dos atributos:
  - Podem conter espaços
  - Podem começar com caractere numérico
  - Podem conter qualquer caractere de pontuação

Atributos

- Exemplo:

```
<livro isbn="85.241.0590-9">
| <capitulo numero="3">
| | <secao>
| | | <para>...</para>
| | | <para>...</para>
| | </secao>
| | <secao>
| | | <para>...</para>
| | </secao>
| </capitulo>
| <capitulo numero="4"> ... ...
</livro>
```

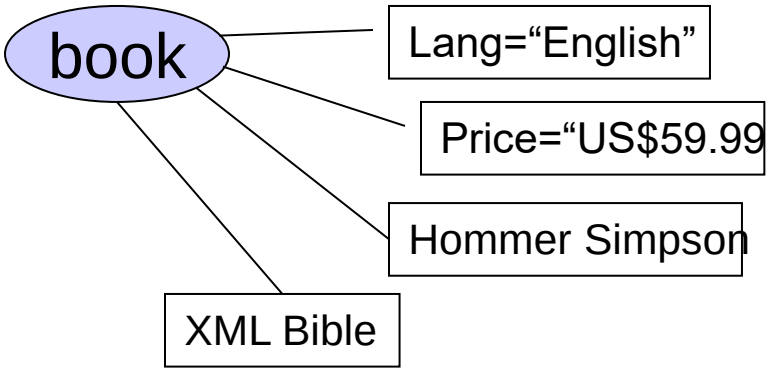
Elementos x Atributos

- Não há regras
- Atributos apresentam algumas restrições
  - Não são extensíveis
  - Não permitem múltiplos valores
  - Não descrevem estruturas
- Recomendação: em geral, preferir elementos, e usar atributos para informações secundárias
- Metadados (dados sobre os dados) devem ser representados como atributos

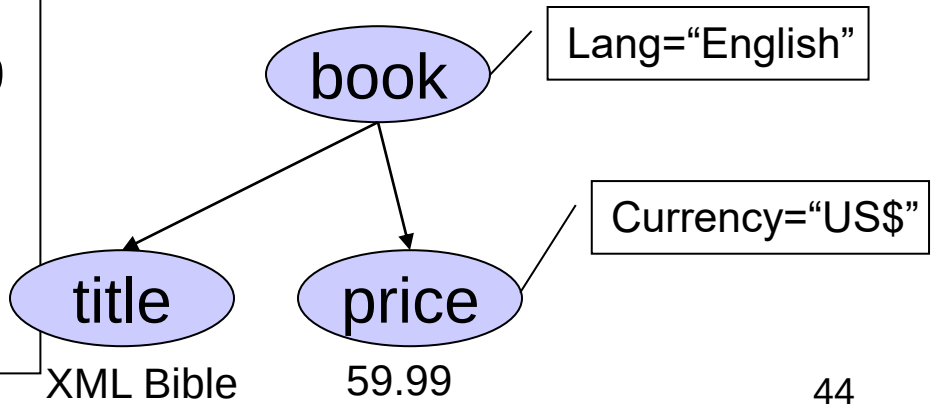
Ex: <price currency="US">59.99</price>

Elementos x Atributos

```
<book  lang="English"
    price="US$59.99"
    title="XML Bible"
    author="Hommer Simpson">
    ...
</book>
```




```
<book lang="English">
  <price currency="US$"> 59.99
  </price>
  <title>XML Bible </title>
        ...
</book>
```



Exemplo de Receituário médico

Ana Maria Marina , 7 anos

Uso interno

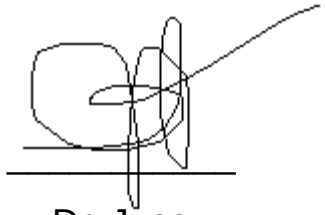
Xarope SemTosse

1 colher 3x ao dia

Uso Externo

Gyellow

aplicar no braço 1x ao dia ao deitar



Dr. Juca

20/10/2001

Exercício

- Representar os dados da receita médica do slide anterior em formato XML
  - Acrescente as marcas aos dados que você criou.
  - Lembre-se: é importante pensar em como estes documentos serão **estruturados**, e não em como serão **apresentados**

Vamos testar?

- Use o Exchanger XML Lite para verificar o documento XML que você criou

ou

- Use o parser RXP... (baixar do site da disciplina)
  - rxp <nome do arquivo XML>

C:_Graca\cursos\jai2000\curitiba\aula-Maior.xml - AT&T Internet Explorer

File Edit View Favorites Tools Help

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!--  ?xml-stylesheet type="text/css" href="aula.css"?  -->
<!--  Apresenta-se o texto, algoritmo e codigo para resolver o problema de escolher
o menor entre tres numeros. Conclui-se com dois testes    -->
- <aulaml id="maior01" prof="Graça" titulo="Verificar o maior entre 3 números">
    <curso cod="sce-180" nome="Introdução à Ciência da Computação" />
    <index termo="maior" />
  + <quadro id="q1" tipo="teoria">
  - <quadro id="q2" tipo="teoria">
    - <texto>
        <paragrafo>Encontra o maior número entre A, B e C</paragrafo>
      - <sequencia>
          <passo>se A > B entao MAIOR = A senao MAIOR = B</passo>
          <passo>se MAIOR < C entao MAIOR = C</passo>
          <passo>fim</passo>
        </sequencia>
      </texto>
    - <codigo>
        <![CDATA[        if A > B then MAIOR = A else MAIOR = B;        ]]>
      </codigo>
    + <teste>
    </quadro>
  </aulaml>
```

C:_Graca\cursos\jai2000\curitiba\aula-Maior.xml - AT&T Internet Explorer

File Edit View Favorites Tools Help

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!--  ?xml-stylesheet type="text/css" href="aula.css"?  -->
<!--  Apresenta                algoritmo e codigo para resolver o problema de escolher
o meno
- <aular                                                                    ">
    <cur
    <ind
  + <qua
  - <qua
    - <te
        <paragrafo>Encontra o maior número entre A, B e C</paragrafo>
      - <sequencia>
          <passo>se A > B entao MAIOR = A senao MAIOR = B</passo>
          <passo>se MAIOR < C entao MAIOR = C</passo>
          <passo>fim</passo>
        </sequencia>
      </texto>
    - <codigo>
        <![CDATA[        if A > B then MAIOR = A else MAIOR = B;        ]]>
      </codigo>
    + <teste>
    </quadro>
  </aulaml>
```

Instruções de Processamento: Mecanismo de inserção de informações explícitas em um documento que são destinadas a alguma aplicação.
*Começa com **<?** e termina com **?>***

C:_Graca\cursos\jai2000\curitiba\aula-Maior.xml - AT&T Internet Explorer

File Edit View Favorites Tools Help

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!--  ?xml-stylesheet type="text/css" href="aula.css"?  -->
<!--  Apresenta-se o texto, algoritmo e codigo para resolver o problema de escolher
o menor entre tres numeros. Conclui-se com dois testes    -->
- <aulaml id="maior01" prof="Graça" titulo="Verificar o maior entre 3 números">
    <curso id="MC-180" nome="Introdução à Ciência da Computação" />
    <index />
    </sequencia>
  </texto>
  - <codigo>
      <![CDATA[        if A > B then MAIOR = A else MAIOR = B;        ]]>
    </codigo>
  + <teste>
  </quadro>
</aulaml>
```

Comentários começam com **<!--** *e terminam com* **-->** *e são ignorados. Não podem acontecer antes da instrução de declaração XML nem dentro de um elemento; não podem conter a seqüência* **--**

C:_Graca\cursos\jai2000\curitiba\aula-Maior.xml - AT&T Internet Explorer

File Edit View Favorites Tools Help

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!--  ?xml-stylesheet type="text/css" href="aula.css"?  -->
<!--  Apresenta-se o texto, algoritmo e codigo para resolver o problema de escolher
o menor entre tres numeros. Conclui-se com dois testes    -->
- <aulaml id="maior01" prof="Graça" titulo="Verificar o maior entre 3 números">
    <curso cod="s
    <index termo=
  + <quadro id="q
  - <quadro id="q
    - <texto>
        <paragrafo
      - <sequencia
          <passo>
          <passo>s
          <passo>fim</passo>
        </sequencia>
      </texto>
    - <codigo>
        <![CDATA[          if A > B then MAIOR = A else MAIOR = B;          ]]>
      </codigo>
    + <teste>
    </quadro>
  </aulaml>
```

Referências a Entidades são marcações que são substituídas com caracteres de dados no processamento do documento. As cinco entidades a seguir são predefinidas por XML:

& < > " '

C:_Graca\cursos\jai2000\curitiba\aula-Maior.xml - AT&T Internet Explorer

File Edit View Favorites Tools Help

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!--  ?xml-stylesheet type="text/css" href="aula.css"?  -->
<!--  Apresenta-se o texto, algoritmo e codigo para resolver o problema de escolher
o menor entre tres numeros. Conclui-se com dois testes    -->
- <aulaml id="maior01" prof="Graça" titulo="Verificar o maior entre 3 números">
    <curso cod="sce-180" nome="Introdução à Ciência da Computação" />
    <index termo="maior" />
  + <quadro id="q1" tipo="teoria">
  - <quadro id="q2" tipo="teoria">
    - <texto>
        <paragrafo>Encontra o maior número entre A, B e C</paragrafo>
      - <sequencia>
          <passo>se A > B entao MAIOR = A senao MAIOR = B</passo>
          <passo>se MAIOR < C entao MAIOR = C</passo>
          <passo>fim</passo>
        </sequencia>
      </texto>
    - <codigo>
        <![CDATA[        if A > B then MAIOR = A else MAIOR = B;        ]]>
      </codigo>
    + <teste>
    </quadro>
  </aulaml>
```

CDATA: todo o texto que aparece entre delimitadores de seção `CDATA` são considerados caracteres de dado:

`<![CDATA[ ... ]]>`

Material de Maria Luiza Machado Campos

C:_Graca\cursos\jai2000\curitiba\aula-Maior.xml - AT&T Internet Explorer

File Edit View Favorites Tools Help

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!--  ?xml-stylesheet type="text/css" href="aula.css"?  -->
<!--  Apresenta-se o texto, algoritmo e codigo para resolver o problema de escolher
o menor entre tres numeros. Conclui-se com dois testes    -->
- <aulaml id="maior01" prof="Graça" titulo="Verificar o maior entre 3 números">
    <curso cod="sce-180" nome="Introdução à Ciência da Computação" />
    <index termo="maior" />
  + <quadro id="q1" tipo="teoria">
  - <quadro id="q2" tipo="teoria">
    - <texto>
        <paragrafo>Encontra o maior número entre A, B e C</paragrafo>
      - <sequencia>
          <passo>se A > B entao MAIOR = A senao MAIOR = B</passo>
          <passo>se MAIOR < C entao MAIOR = C</passo>
          <passo>fim</passo>
        </sequencia>
      </texto>
    - <codigo>
```

```
<passo>se A &gt; B
       entao MAIOR = A senao MAIOR = B
</passo>
<passo>se MAIOR &lt; C entao MAIOR = C
</passo>
```