



XML & Cia: Introdução

Renata Pontin de Mattos Fortes

Instituto de Ciências Matemáticas e de Computação

Universidade de São Paulo

{renata}@icmc.usp.br

Agradecimento especial



Prof.Dr.Maria da Graça Pimentel

pelo material cedido

Roteiro

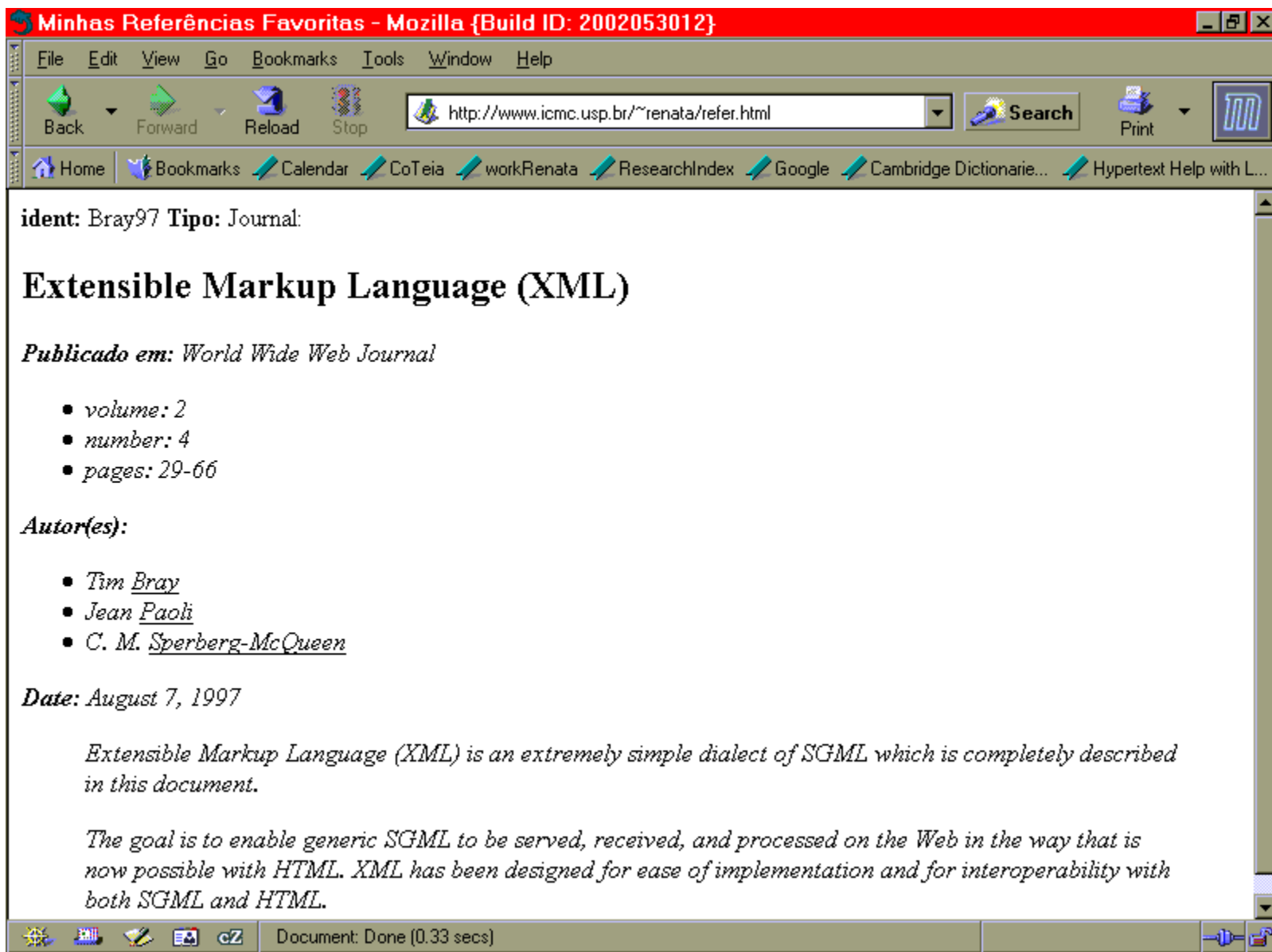


- **Introdução: XML**
- Sintaxe e validação: DTD e XML Schema
- Transformação e apresentação: XSLT e XPath

HTML



- HyperText Markup Language
 - Uma linguagem para a especificação da apresentação e estruturação de documentos em browsers
 - HTML não fornece suporte a aplicações flexíveis e interoperáveis



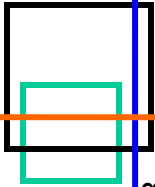
```

<html>
<head><title>Minhas Referências Favoritas</title></head>
<body>
<b>ident: </b>Bray97 <b>Tipo: </b>Journal:<p>
<h2>Extensible Markup Language (XML)</h2><p>
<i><b>Publicado em:</b> World Wide Web Journal
<ul><li> volume: 2</li><li> number: 4</li>
<li> pages: 29-66</li></ul>
<b>Autor(es):</b>
<ul><li> Tim <u>Bray</u></li><li> Jean <u>Paoli</u></li>
<li> C. M. <u>Sperberg-McQueen</u></li></ul>
<b>Date: </b>August 7, 1997<br>
<blockquote><i>Extensible Markup Language (XML)</i> is
an extremely simple dialect of <i>SGML</i> which is
completely described in this document.
<p></p>
The goal is to enable generic SGML to be served,
received, and processed on the Web in the way
that is now possible with <i>HTML.</i> XML has
been designed for ease of implementation and for
interoperability with both SGML and HTML.
</html>

```

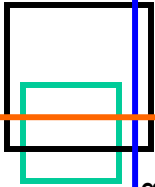
Limitações de HTML

- HTML é direcionado à apresentação de uma classe particular de documentos
 - Títulos, cabeçalhos, tabelas, listas, etc.
 - `<h2>` Extensible Markup Language (XML)
`</h2>`
 - ``Tim `<u>`Bray`</u>` ``



Limitações de HTML

- HTML não é extensível
 - Uma aplicação não pode definir novos elementos e tê-los reconhecidos por outras aplicações
 - <author>
 - <date>



Limitações de HTML

- Um documento HTML não é reutilizável
 - Não é possível gerar automaticamente um novo documento a partir de um documento HTML
 - Na referência bibliográfica em HTML, por exemplo, um novo documento com apenas título e resumo

Limitações de HTML



- Um documento HTML corresponde a uma visão particular da informação
 - A única visão que existe é a da aplicação que gerou o documento HTML

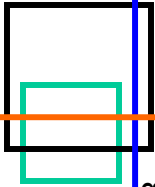
Limitações de HTML



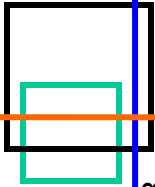
- Pouca, ou quase nenhuma semântica pode ser extraída de um documento HTML
 - Na referência bibliográfica em HTML, por exemplo, não é possível isolar a parte que corresponde ao nome do autor do restante do documento

- ***Standard Generalized Markup Language***

- Uma meta-linguagem para a especificação da estrutura de documentos
- Exemplo
 - Referência bibliográfica

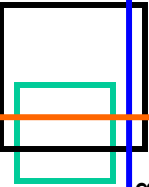


```
<bibref id="Bray97" type="Journal">
<title>Extensible Markup Language (XML)</title>
<publication volume="2" number="4" pages="29-66">
World Wide Web Journal</publication>
<authors>
<name>Tim <lastname>Bray</lastname></name>
<name>Jean <lastname>Paoli</lastname></name>
<name>C. M. <lastname>Sperberg-McQueen</lastname></name>
</authors>
<date>August 7, 1997</date>
<ul><li> C. M. <u>Sperberg-McQueen</u></li></ul>
<abstract>
<enfa>Extensible Markup Language (XML)</enfa> is
an extremely simple dialect of <enfa>SGML</enfa> which
is completely described in this document.<sep/>
The goal is to enable generic SGML to be served,
received, and processed on the Web in the way
that is now possible with <enfa>HTML.</enfa> XML has
been designed for ease of implementation and for
interoperability with both SGML and HTML.
</abstract>
</bibref>
```



Vantagens de SGML

- Estrutura hierárquica
 - SGML permite a representação da estrutura hierárquica dos elementos em um documento
 - O elemento <bibref> contém os elementos <title>, <publication>, <authors>, <date> e <abstract>

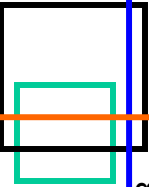


Vantagens de SGML

- Flexibilidade

- SGML não dita quais tipos de elementos devem ser criados ou como eles se relacionam, isto fica a critério do autor da aplicação

- Excluir o elemento <abstract>
- Modificar o elemento <date> para incluir <dd>, <mm> e <yy>
- Extrair os elementos <title> e <publication>



Vantagens de SGML

- Especificação formal
 - A estrutura e o tipo dos elementos contidos em um documento SGML são formalmente definidos por uma gramática
 - A aplicação pode validar o documento através da sua definição
 - Aplicações genéricas para análise léxica e sintática (*parsers*) podem ser construídas

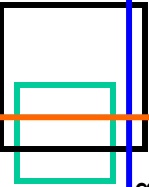
Vantagens de SGML



- Reusabilidade
 - Tanto definições dos documentos quanto os documentos propriamente ditos podem ser utilizados por diversas aplicações
 - A partir de um conjunto de documentos como o da referência bibliográfica em SGML, é possível produzir um novo documento de todos os artigos publicados por “Tim Bray”

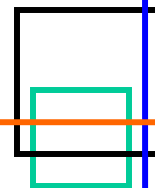
Vantagens de SGML

- Representação legível
 - Um documento SGML tem conteúdo textual e pode ser editado em qualquer editor e ser lido e compreendido com facilidade pelo usuário
 - `<date>August 7, 1997</date>`



SGML *versus* HTML

- HTML
 - Todas as limitações de HTML são superadas com o uso de SGML
- SGML
 - O custo do processamento de documentos SGML impede seu uso direto em aplicações rodando sobre o ambiente distribuído da Web



XML: Linguagem de Marcação Extensível

- W3C: World Wide Web Consortium

- <http://www.w3.org>



Quem?

- Extensible Markup Language – XML

- Recomendação W3C, 10 de Fevereiro, 1998, <http://www.w3.org/TR/REC-xml/>

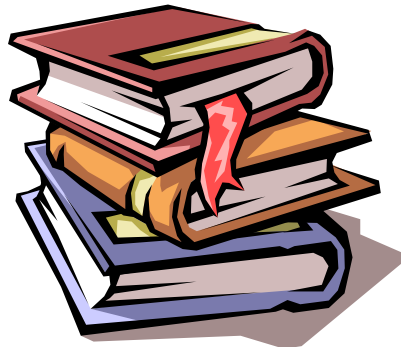
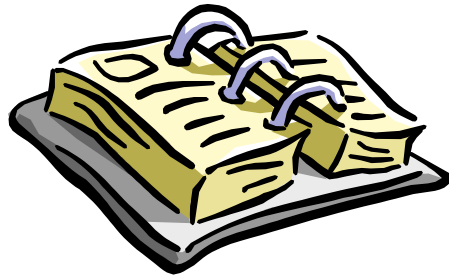
Quando?

Como?

Onde?

Por quê?

O que é um documento???



Capítulo 3

Papel de Padrões XML em OHSs

3.1 Considerações Iniciais

Ao abstrair a entidade "documento", pode-se imaginá-la sob as seguintes dimensões: conteúdo, estrutura, apresentação, semântica, metadados¹ e malha de hipertexto.² Um documento XML representa, basicamente, duas dimensões de informação: o conteúdo propriamente dito e a estrutura organizacional deste conteúdo. A Recomendação XML (Bray et al., 1998) possibilitou a definição de especificações que representam as demais dimensões de informação de um documento.

Paralelamente ao trabalho do W3C (World Wide Web Consortium), desenvolveram-se várias pesquisas na comunidade de hipertexto aberta (OHSWG, 2001) que visavam entender a infra-estrutura da Web com serviços que armazenam as informações de ligação externamente ao conteúdo dos documentos. Para Carr and Hall (1998), OHSs como o DLS Microcom e o DHM fizeram surgir questões relevantes para a comunidade de hipertexto aberta sobre serviços de ligação, como:

- Definição dos tipos de interface com o usuário para a apresentação de ligações;
- Variedade de comportamentos relacionados a ligações;
- Semântica associada a conjuntos de ligações que integram hiperdocumentos, etc.

¹Por exemplo, metadados usam como metadados: autor, título, assunto, editora e a localização física de cada livro.

²Identificam a numeração, índices de tabelas e figuras, notas de rodapé e bibliografia podem ser consideradas como malha de hipertexto sobre documentos, em geral.



Capítulo 3

Papel de Padrões XML em OHSs

3.1 Considerações Iniciais

Ao abstrair a entidade “documento”, pode-se imaginá-la sob as seguintes dimensões: conteúdo, estrutura, apresentação, semântica, metadados¹ e malha de hipertexto.² Um documento XML representa, basicamente, duas dimensões de informação: o conteúdo propriamente dito e a estrutura organizacional deste conteúdo. A Recomendação XML (Bray et al., 1998) possibilitou a definição de especificações que representam as demais dimensões de informação de um documento.

Paralelamente ao trabalho do W3C (*World Wide Web Consortium*), desenvolveram-se várias pesquisas na comunidade de hipernúdia aberta (OHSWG, 2001) que visavam estender a infra-estrutura da Web com serviços que armazenam as informações de ligação externamente ao conteúdo dos documentos. Para Carr and Hall (1998), OHSs como o DLS Microcosm e o DHM fizeram surgir questões relevantes para a comunidade de hipernúdia aberta sobre serviços de ligação, como:

- Definição dos tipos de interface com o usuário para a apresentação de ligações;
- Variedade de comportamentos relacionados a ligações; e
- Semântica associada a conjuntos de ligações que interligam hiperdocumentos, etc.

¹Por exemplo, bibliotecas usam como metadados: autor, título, assunto, editora e a localização física de cada livro.

²Referências a sumário, índices de tabelas e figuras, notas de rodapé e bibliografia podem ser consideradas como malha de hipertexto sobre documentos, em geral.

- Conteúdo
- Estrutura
- Apresentação
- Semântica
- Metadados
- Hipertexto

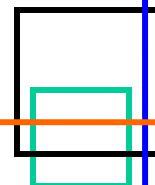
E um documento da Web???

```
<html>
<head>
  <title>Exemplo de HTML</title>
  <meta name="GENERATOR"
    content="Microsoft FrontPage
    4.0">
</head>
<body>
  <i><h1>cabeçalho</h1></i>
  <p>parágrafo<b>texto em
    negrito</b>
  <p>texto em novo parágrafo</p>
  <a href =
    "http://www.w3.org">W3C</a>
</body>
</html>
```

- Simples 😊
- Fácil aprendizado 😊
- Checagem de sintaxe 😐
- Extensibilidade 😞
- Consciência de conteúdo 😞
- Carência de semântica 😞
- Visa à apresentação 😞
- Intercâmbio de dados 😞
- Conteúdo não-computável 😞
- Pouco reuso 😞

Os outros documentos???

- Menus de restaurantes
- Bilhetes de teatro
- Documentos financeiros
- Gráficos
- Curriculum vitae
- Histórico escolar
- Monografia de conclusão
- Protocolo de comunicação
- Notícias jornalísticas
- Fórmulas matemáticas
- Fórmulas químicas
- Programas de computador
- Apólices de seguro
- Multimídia
- Modelo de ER
- Catálogo de livros



Demanda atual de aplicações Web

- Extensibilidade do conjunto de marcadores que
 - ... permite que autores ou comunidades criem seus **próprios marcadores** para melhor definirem seus documentos de interesse
 - ... permite que aplicações possam associar **significado** a dados e campos do documento o que viabilizaria o **processamento** automático dos documentos
 - ... permite a construção de aplicações mais apropriadas para dispositivos **portáteis** e de poucos recursos, por exemplo

Por que?

XML: Linguagem de Marcação Extensível

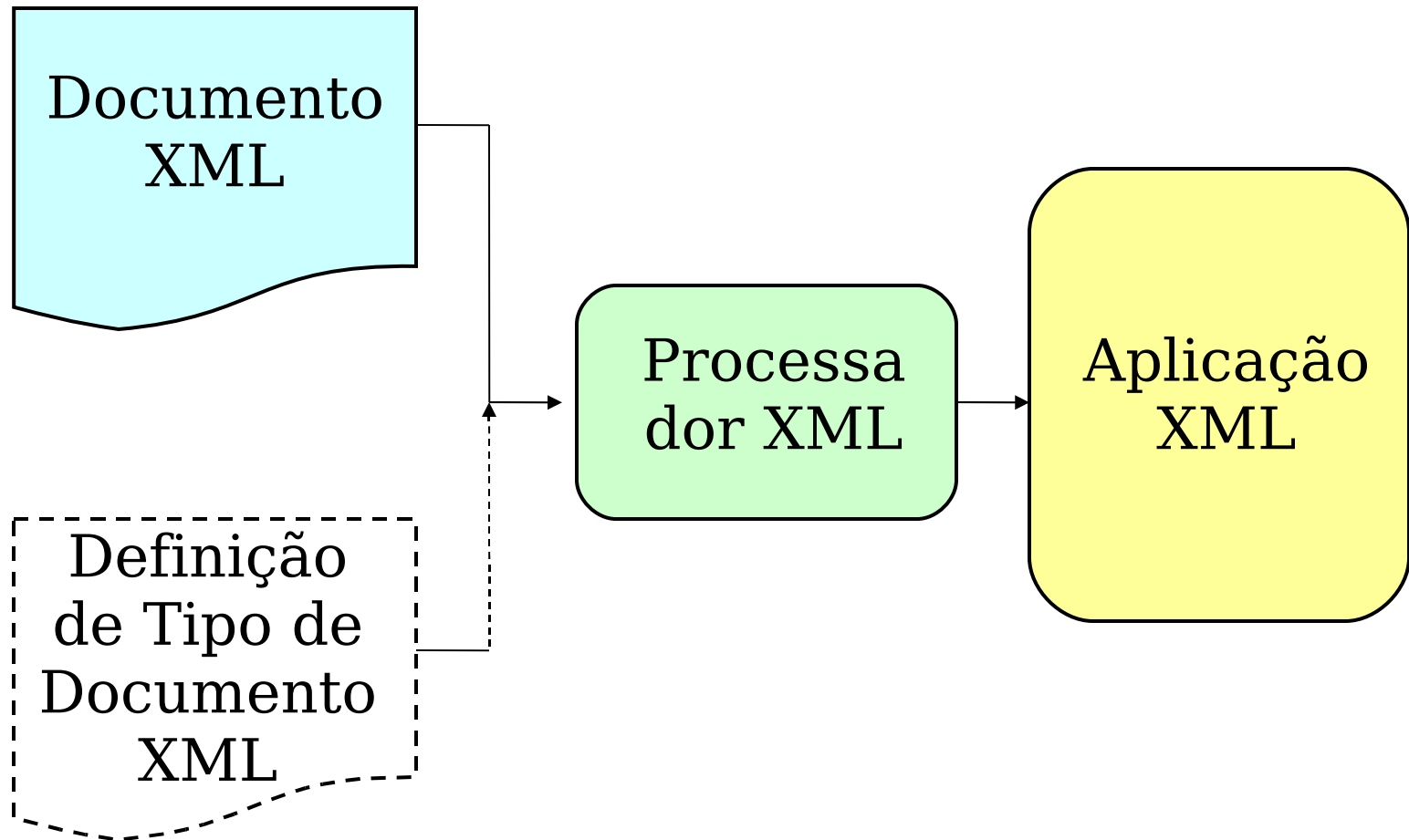
- É uma linguagem de marcação apropriada à **representação** de dados, de documentos e demais entidades cuja essência se fundamenta na capacidade de **agregar** informações
- XML é uma **linguagem** ao estabelecer **regras gerais** às quais documentos em **conformidade** com XML devem respeitar

Objetivo Principal de XML



- Permitir a especificação de documentos abertos apropriados para intercâmbio entre aplicações na Web

O objetivo principal de XML



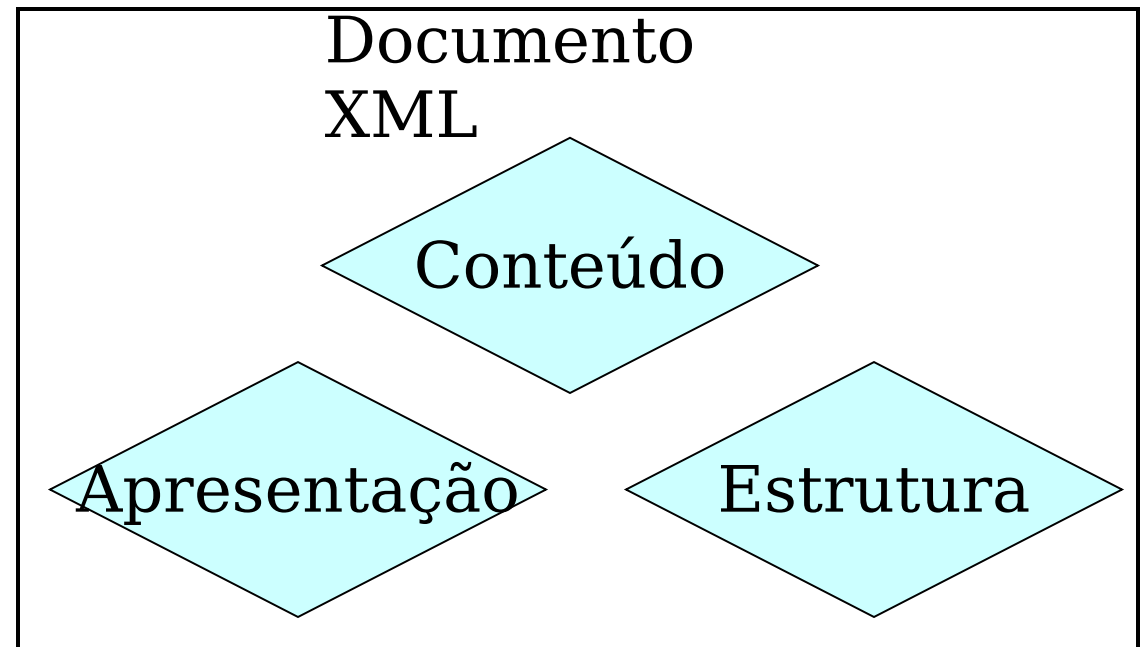
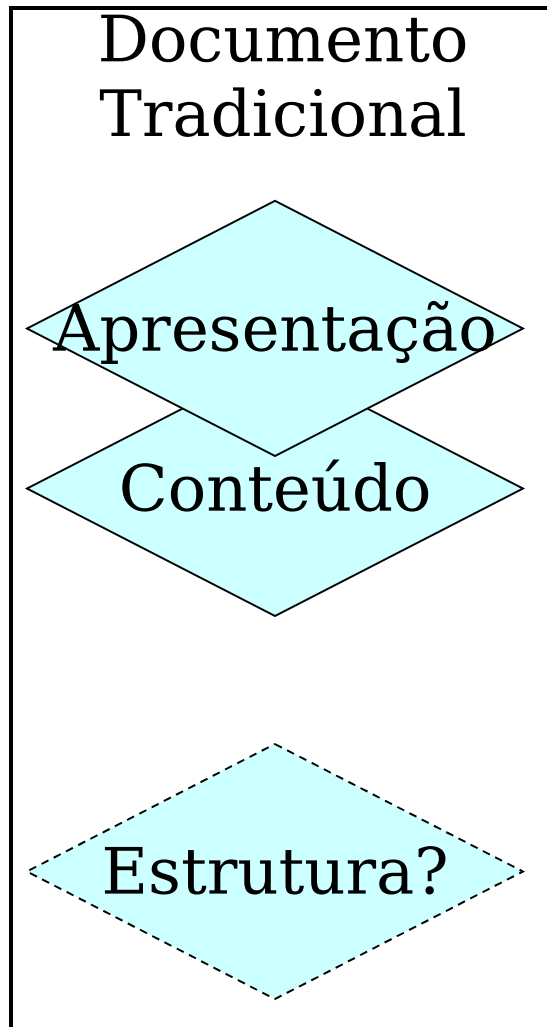
Dimensões de um documento XML

- Conteúdo

- Estrutura

```
<slides>
  <slide name="introdução">
    <p>Introdução AQUI</p>
  </slide>
  <slide name="background">
    <p> background AQUI</p>
  </slide>
  <slide name="resultados">
    <p> resultados AQUI</p>
  </slide>
  <slide name="conclusão">
    <p>Conclusão AQUI</p>
  </slide>
</slides>
```

Dimensões de um documento XML



Dimensões de um documento XML



- Outras dimensões de um documento XML
 - **Estrutura** e **Semântica**: DTD e Esquema XML
 - **Apresentação**: CSS e XSL (XPath, XSLT e XSL-FO)
 - **Metadados** e mais **semântica**: RDF e Esquema RDF
 - Estrutura de **hipertexto**: XLink e XPointer
- Processamento de documentos XML
 - Parsers, DOM, SAX, aplicações ..
- <http://www.w3.org>

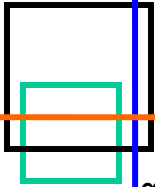
XML: linguagem ou meta-linguagem???



- XML é uma **linguagem**:
 - estabelece regras gerais às quais documentos em conformidade com XML devem respeitar
- XML é também uma **meta-linguagem**:
 - provê recursos para a definição de gramáticas que caracterizam linguagens para classes de documentos específicos com conjunto de elementos, atributos e regras de composição bem determinados

Vocabulários XML

- RDF e RDF Schema
- DAML-OIL
- XML-EDI
- OFX
- CML
- MathML
- XLink
- XPointer
- XPath
- XSLT
- XSL-FO
- DOM
- SAX
- WebDAV
- SOAP
- XML-RPC
- SVG
- VoiceML
- CDF
- SMIL
- ...




XML como meta-linguagem ...



- **CML**: Linguagem de Marcação Química
 - descrição de fórmulas químicas
- **OFX**: Intercâmbio Financeiro Aberto
 - troca de faturas, recibos, extratos ...
- **WML**: Linguagem de Marcação Sem Fio
 - para dispositivos móveis, como celulares e handhelds
- **MathML**: Linguagem de Marcação Matemática
 - descrição de sentenças e fórmulas matemáticas
- **SVG**: Gráficos Vetoriais Escaláveis
 - ideal para especificação de gráficos vetoriais

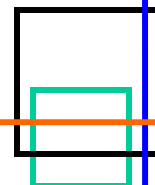
XML como meta-linguagem ...

- **Esquema XML:** gramática XML com forte tipagem de dados
- **XSLT:** transforma um documento XML em outro documento (XML, HTML, RTF, PS, PDF, ...)
- **RDF:** Framework para Descrição de Recursos
 - descrição de recursos por metadados e **semântica**
- **XHTML:** HTML segundo as regras da especificação XML 
- **XML-RPC:** Chamada Remota a Procedimentos em XML
 - protocolo de comunicação em ambientes distribuídos
- **SOAP:** Protocolo de Acesso Simples a Objetos
 - protocolo de comunicação em ambientes distribuídos
- **SMIL:** Linguagem de Integração de Multimídia Sincronizada
 - especificação de apresentações multimídia

Sintaxe e Validação XML

Documentos XML

- Conceitos básicos
 - Documentos XML são textos formados por caracteres do conjunto Unicode, contendo caracteres de dado e informações de marcação explicitamente separados
 - Informações de marcação
 - Comentários, referências a caracteres e entidades, delimitadores de seção CDATA, elementos, instruções de processamento e Definições de Tipo de Documento (DTDs)



Declarações

Instrução de
Processamento

```
<?xml version="1.0" encoding="ISS0-8859-1" ?>
<!DOCTYPE thesis [
  <!ENTITY chapter1 SYSTEM "chap1.xml">
  <!ENTITY chapter7 SYSTEM "chap7.xml">
  <!ELEMENT thesis (abstract,chapter+)>
  <!ATTLIST thesis author CDATA #REQUIRED">
  <!ELEMENT abstract (#PCDATA)>
  <!ELEMENT chapter (title,body)>
  <!ELEMENT title (#PCDATA)>
  <!ELEMENT body (#PCDATA)>
]>
```

DTD

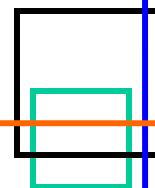
```
<thesis>
<abstract>Nesta tese apresentamos ...</abstract>
&chapter1;
<!-- falta pouco ...-->
&chapter7;
</thesis>
```

Comentários

Elementos

Conjuntos de caracteres aceitos

- ASCII
 - 1 byte, 7 bits → 128 combinações
- ISO 8859-1 Latin-1
 - 1 byte, 8 bits → 256 combinações (ASCII + caracteres para maioria das línguas da Europa Ocidental – **inclusive Português**)
- ISO 8859-(2...15)
 - 1 byte, 8 bits → 256 combinações (ASCII + caracteres para outros conjuntos de línguas)



Unicode: conjunto **padrão** para XML

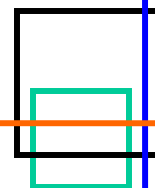
- 2 bytes → 65534 combinações
- Atualmente pouco mais de 40000 utilizadas
- 0-127 Latin Básico – ASCII (inglês USA)
- 126-255 Latin-1 **português**, inglês UK, francês, italiano, espanhol, etc.
- ...
- 19966-40959 Ideogramas – chinês, japonês, etc.
- 57344-63743 Uso privado por desenvolvedores

UTF-8

- Versão compacta do Unicode
- Utiliza apenas 1 byte p/ maioria dos caracteres
- ... ao custo de usar 3 p/ os menos comuns
- p/ inglês → redução de +- 50%

<?XML version="1.0" encoding="UTF-8"?>

Regras para documentos XML



- Nomes dos elementos são sensíveis à caixa

<mensagem> ≠ <Mensagem> ≠ <MENSAGEM>

- Nomes devem começar com letra ou com _ ; o restante pode incluir letras, dígitos, hifens ou o caracter _ , mas **nunca** o caracter de espaço

<slide_1> 😊 **<1slide>** ☹️ **<slide1>** 😊 **<sli de1>** ☹️

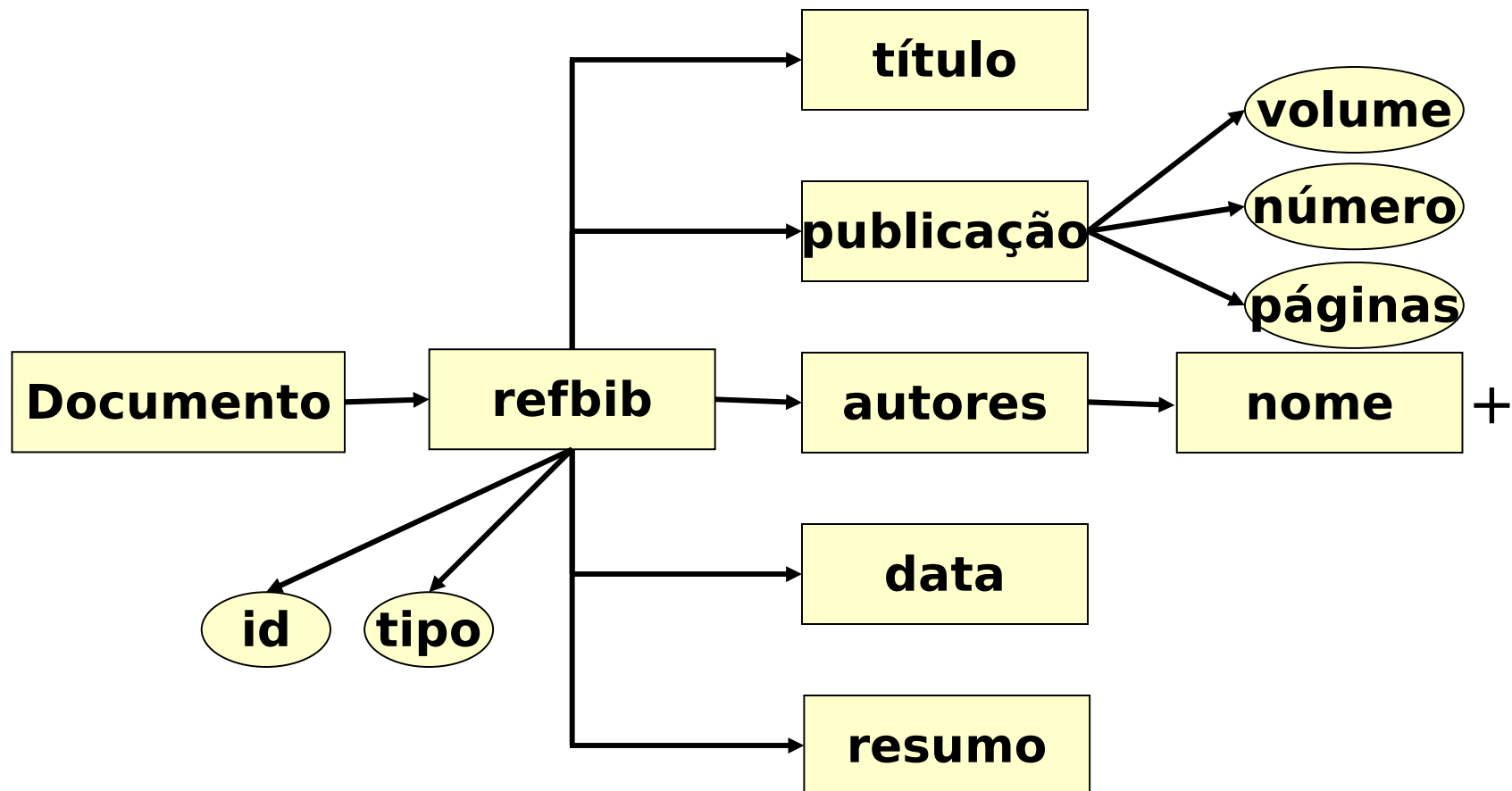
- Utilizar atributo **encoding** em instruções de processamento

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

Estrutura física de documentos XML

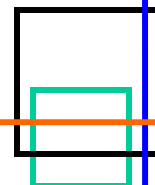
```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE bibref SYSTEM "bibref.dtd" >
<refbib id="Bul" tipo="Periódico">
  <título>Padrões XML e Sistemas Hipermídia Abertos</título>
  <publicação volume="2" número="4" páginas="29-66">
    Cadernos de Computação</publicação>
  <autores>
    <nome>R.F. Bulcão Neto</nome>
    <nome>M.G.C. Pimentel</nome>
  </autores>
  <data>07 de Outubro de 2002</data>
  <resumo>
    A linguagem XML tem papel fundamental nas pesquisas em
    SHAs ...
  </resumo>
</refbib>
```

Estrutura lógica de documentos XML



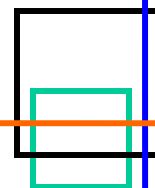
Documentos XML

- Documento bem-formado
 - Documento que segue as definições léxicas e sintáticas de XML
 - Uma das condições para que um documento seja bem-formado é que as estruturas lógica e física estejam **aninhadas**

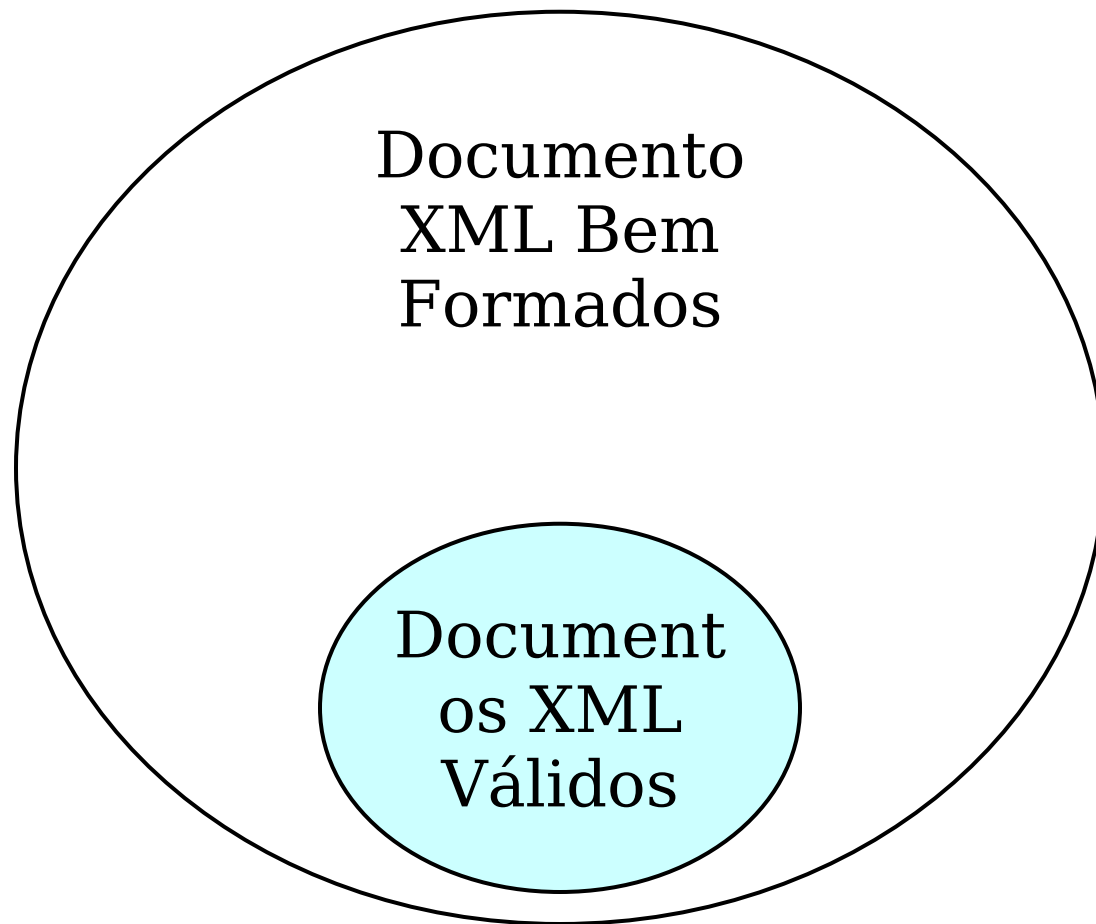


Documentos XML

- Documento válido
 - Documento que segue gramática estabelecida por uma *Document Type Definition* (DTD)

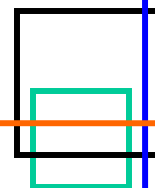


Documentos XML



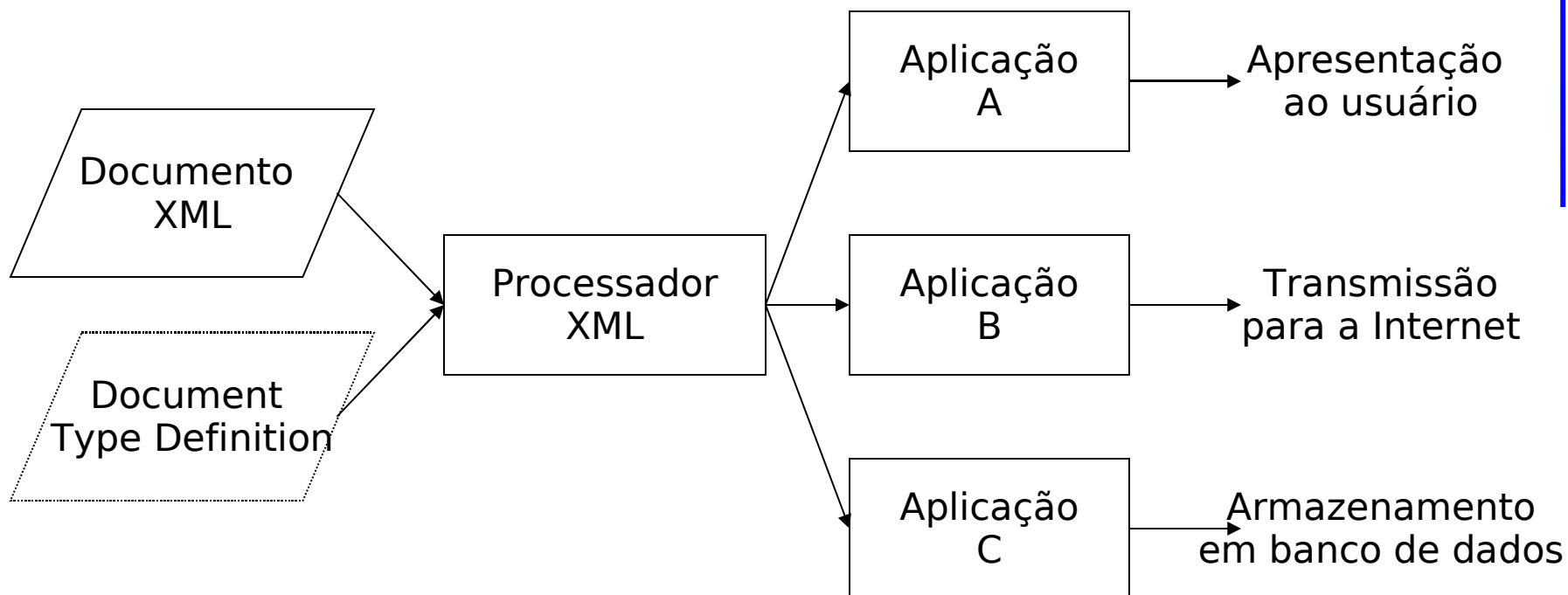
Documentos XML

- Uma aplicação que utilize documentos XML deve **processar** os documentos e **verificar** se seu conteúdo está de acordo com as regras de formação de um documento XML em geral (documento bem formado) e, se for o caso **validar sua estrutura** e conteúdo frente à gramática correspondente, definida na **DTD** (documento válido).



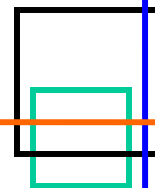
Documentos XML

■ Processador XML



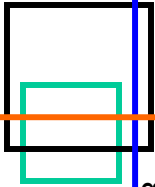
Documentos XML

- Processador XML
 - Processadores Não-validadores
 - Verificam se um documento é “bem-formado” (se está de acordo com as regras gerais de XML que se aplicam a quaisquer documentos)
 - Processadores Validadores
 - Verificam se um documento é bem formado e se o documento está em conformidade com a DTD que define sua gramática



Notação e Especificação XML

- Notação básica e estruturas lógicas
 - Um documento XML é constituído de marcações (*markup*) e de *character data*
 - *Markup*: *start-tags*, *end-tags*, *emph-element tags*, referências a entidades, comentários, seções *CDATA*, declarações de tipo de documento e instruções de processamento
 - *Character data*: todo texto que não é marcação constitui o *character data* do documento



Notação e Especificação XML

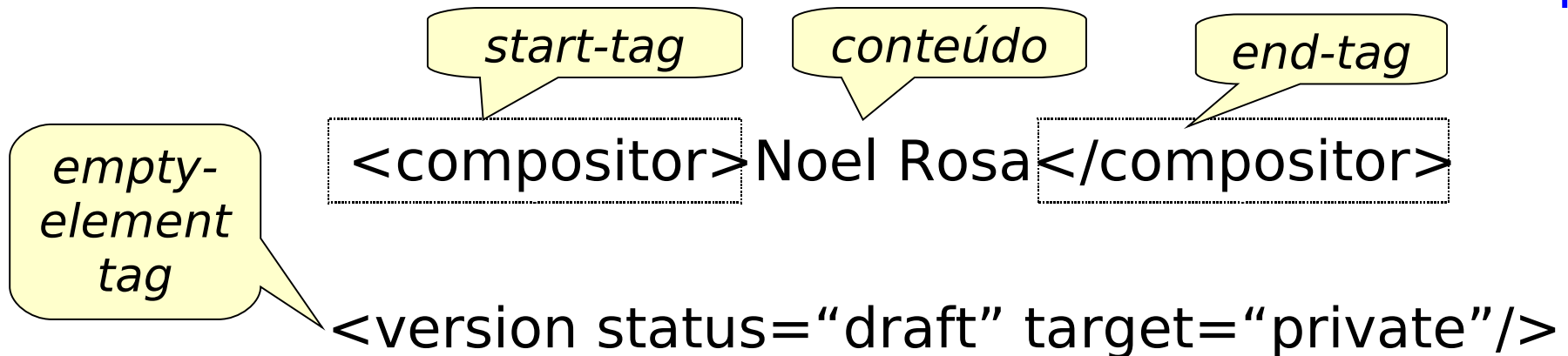
■ Elemento

- Construtor básico de um documento.
- De maneira geral, um elemento é qualquer cadeia de caracteres que aparece entre os caracteres delimitadores < e >, desde que não esteja contido em um comentário ou em uma seção CDATA.
- Pode conter outros elementos, outras marcações (comment, PI, entity references, etc)

Notação e Especificação XML

- Elemento

- Cada elemento tem um tipo, identificado por um nome, e pode ter um conjunto de especificações de atributos associados



Notação e Especificação XML

- Elementos XML são extensíveis.
 - Documentos XML podem ser estendidos para carregarem mais informações.

```
<note>  
  <to> Tover </to>  
  <from> Jani </from>  
  <body> Don't forget me this weekend</body>  
  <heading> Reminder </heading>  
</note>
```

Notação e Especificação XML

- Elementos XML possuem relacionamentos

- Elementos são relacionados como pais e filhos.

Book é o elemento pai de title, prod e chapter

book é o elemento raiz

title, prod e chapter são elementos irmãos

<book>

<title> My first XML </title>

<prod id="123" midia="paper"/>

<chapter> Introduction to XML

<para> What is XML? </para>

</chapter>

</book>

title, prod e chapter são os elementos filho de book

Notação e Especificação XML

- Elementos XML possuem conteúdo.
 - Elementos podem ter diferentes tipos de conteúdo.

```
<book>
```

```
  <title> My first XML </title>
```

```
  <prod id="123" midia="paper"/>
```

```
  <chapter> Introduction to XML
```

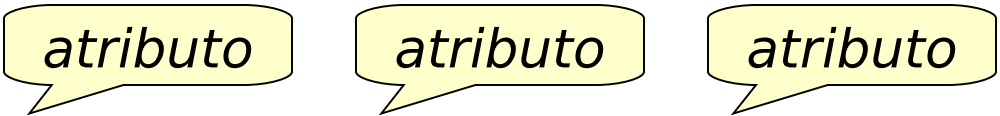
```
    <para> What is XML? </para>
```

```
  </chapter>
```

```
</book>
```


Notação e Especificação XML

- Atributo
 - Um atributo é um par (nome, valor) presente na start-tag do elemento, logo após seu nome


<publication volume="2" number="4" pages="29-66">
World Wide Web Journal</publication>

Os valores dos atributos devem estar entre aspas.
Um atributo não pode aparecer mais de uma vez no mesmo elemento.

Notação e Especificação XML

■ Entidades

- Uma referência a uma entidade é da forma *&nome da entidade*
- Entidade externa: &chapter1;
- Entidade interna:

Graça representa Graça

x < y representa $x < y$

Notação e Especificação XML

- CDATA

- Uma seção CDATA permite a inclusão de trechos que devem ser interpretados como caracteres e não como elementos de marcação

```
<![CDATA[ *p = &q; b = (i <= 3); ]]>
```

Notação e Especificação XML

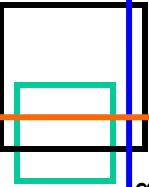
- Comentários
 - Não fazem parte do conteúdo do documento



Notação e Especificação XML

- Declaração XML
 - Documentos XML podem, e devem, começar com uma declaração XML

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



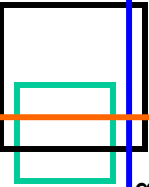
Notação e Especificação XML

- Instruções de processamento
 - Os parsers XML apenas repassam essas informações para a aplicação



Notação e Especificação XML

- Documento bem formado
 - Se presente, a instrução de processamento deve literalmente iniciar o documento
 - Todo documento deve ter um elemento (raiz) que inclui todos os demais
 - Não é permitido desrespeitar a estrutura de aninhamento de elementos



fim

- **Sintaxe e validação: DTD e XML Schema**
- Transformação e apresentação: XSLT e XPath