

UFSC-CTC-INE

Curso de Sistemas de Informação

INE 5600 – Bancos de Dados III

Introdução a Dados na Web e XML

Por quê o interesse por dados na Web?

- Web: maior fonte de dados públicos em diversos domínios
 - páginas HTML, bases de dados, ...
- Dados úteis para consumo humano
 - busca por informação em domínios específicos
 - complementação / cruzamento / integração / ... de dados, ...
- Desafios
 - alta heterogeneidade de representação
 - inexistência de esquema
 - carência de mecanismos eficientes de busca

Natureza dos Dados na Web


- Dados Estruturados
 - informação com padronização de apresentação
 - atributos explícitos ou não
 - exemplos:
 - *Web tables*
 - *Web lists*
 - *Web records*
 - *Deep Web*
 - ...

Web Tables

Activity	Description	Venue	Day	Time	A2L Price *	Full Price
Sh'bam	Fun loving, carefree dance based class with great music, disco lights and a party atmosphere.	Beach Leisure Centre	Mon	1:15 – 2:00pm	£3.00	£5.95
Badminton	A badminton coaching session for teenagers and adults with a learning disability. This session encourages fun, fitness and friendship. (During school term only)	Cults Sports Complex, at Cults Academy	Mon	7:00 - 8:30pm	£2.55	£5.10
Gym session	Help and advice in the gym from instructors who will support, encourage and advise. One to one inductions with an instructor can also be arranged.	Beach Leisure Centre	Tues	10:00am – 12:00pm	£3.75	£7.40
Swimming	Dedicated public sessions for swimmers with a disability and their family, friends or carers. (includes use of small flume as standard at the Beach Leisure Centre only)	Bridge of Don Pool	Wed	1:00 – 2:00pm	FREE	£4.30
		Bucksburn Pool	Thu	1:30 – 2:30pm	FREE	£4.30
		Tullos	Thu	1:00 - 2:00pm	FREE	£4.30
		Beach Leisure Centre	Sat	5:00 – 6:00pm	£4.30	£4.30
Boccia	Fun session for teenagers and adults to learn to play boccia through skills, activities and game play. (during school term only)	Westburn Park	Wed	6:30 - 7:30pm	£2.55	£5.10
Ice Skating	Inclusive public skating lesson	Linx Ice Arena	Fri	10:00 – 12:00pm	£3.20 + £0.95 skate	£6.35 + £1.85

No.	Portrait	President	Term of office	Party	Term [a]	Previous office	Vice President
1		George Washington February 22, 1732 – December 14, 1799 (aged 67) [11][12][13]	April 30, 1789 [b] – March 4, 1797	Non-partisan [14]	1 (1789)	Commander-in-Chief of the Continental Army (1775–1783)	John Adams
					2 (1792)		
2		John Adams October 30, 1735 – July 4, 1826 (aged 90) [15][16][17]	March 4, 1797 – March 4, 1801 [c]	Federalist	3 (1796)	1st Vice President of the United States	Thomas Jefferson

Deep Web



New

Used

Certified Pre-Owned

All

What's this?

Ford

All Models

F150

F250

F350

F450

Fiesta

Flex


Focus

Focus Electric


Focus RS

Focus ST


Fusion



A SMALL BUSINESS



Driving Smart



New 2013 Fusion


\$28,360

New

Black, 4 door, FWD, Sedan,
Gas I4 2.5L/152, Stock# 13F470.

Desoto Dodge Chrysler Jeep ~ 25 mi. away
877-364-8584 [Email Dealer](#)

☐ Save/Compare



Used 2012 Fusion


\$27,888

14 mi.

Ginger, 4 door, FWD, Sedan,
1-Speed Continuously Variable Ratio, Gas/Electric
I4 2.5L/152, Stock# P5651.

Desoto Dodge Chrysler Jeep ~ 25 mi. away
888-828-6058 [Email Dealer](#)

☐ Save/Compare ☒ Free CARFAX Report



Used 2012 Fusion


\$22,888

10,054 mi.

Black, 4 door, FWD, Sedan,
6-Speed Automatic w/SelectShift, Gas V6
3.5L/213, Stock# P5728.

Desoto Dodge Chrysler Jeep ~ 25 mi. away
888-828-6058 [Email Dealer](#)

☐ Save/Compare ☒ Free CARFAX Report



Used 2013 Fusion

\$22,489

5,966 mi.

Gray, 4 door, FWD, Sedan,
6-Speed Automatic, 2.5L I4 16V MPFI DOHC,
Stock# DR230294.

Matthews-Currie ~ 21 mi. away
888-362-5796 [Email Dealer](#)

Natureza dos Dados na Web

- Dados Não-Estruturados
 - dados de mídias não-textuais
 - metadados podem estar disponíveis junto aos arquivos de dados
 - exemplos:
 - imagens
 - áudios
 - vídeos
 - ...

Natureza dos Dados na Web

- Dados Semiestruturados
 - dados com alguma estrutura (textual) explícita
 - parte não-estruturada composta por diferentes mídias (texto, imagem, ...)
 - exemplos:
 - páginas HTML de modo geral
 - documentos (e-mails, ebooks, ...)
 - ...

Dados Semiestruturados

[Dbworld] BigDat 2017: early registration 26 August



De **Carlos Martin**
 Remetente **Dbworld**
 Para **dbworld@cs.wisc.edu**
 Responder para **dbworld_owner@yahoo.com**
 Data **Sex. 02:13**

3rd INTERNATIONAL WINTER SCHOOL ON BIG DATA

BigDat 2017

Bari, Italy

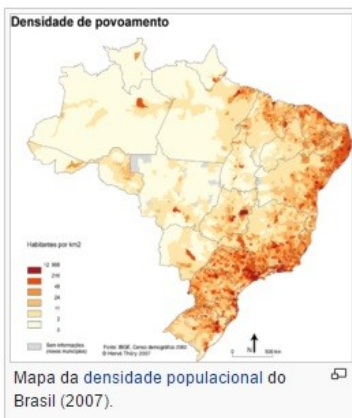
February 13-17, 2017

Organized by:
 University of Bari "Ald
 Rovira i Virgili Univer

<http://grammars.grlmc.c>

--- Early registration

AIM:





A população do Brasil, conforme registrado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em 2010, foi de 190 755 799 habitantes^[167] (22,43 habitantes por quilômetro quadrado),^[168] com uma proporção de homens e mulheres de 0,96:1^[169] e 84,36% da população definida como urbana.^[4] A população está fortemente concentrada nas regiões Sudeste (80,3 milhões de habitantes), Nordeste (53,1 milhões de habitantes) e Sul (27,4 milhões de habitantes), enquanto as duas regiões mais extensas, o Centro-Oeste e o Norte, que formam 64,12% do território brasileiro, contam com um total de apenas trinta milhões de habitantes.^[4]

A população brasileira aumentou significativamente entre 1940 e 1970, devido a um declínio na taxa de mortalidade, embora a taxa de natalidade também tenha passado por um ligeiro declínio no período. Na década de 1940 a taxa de crescimento anual da população foi de 2,4%, subindo para 3% em 1950 e permanecendo em 2,9% em 1960, com a expectativa de vida subindo de 44 para 54 anos^[170] e para 73,9 anos em 2013.^[171] A taxa de aumento populacional tem vindo a diminuir desde 1960, de 3,04% ao ano entre 1950–1960 para 1,05% em 2008 e deverá cair para um valor negativo, de -0,29%, em 2050,^[172] completando assim a transição demográfica.^[173]

Os maiores aglomerados urbanos do Brasil são as áreas metropolitanas de São Paulo (com 21 090 792 habitantes), Rio de Janeiro (12 280 702) e Belo Horizonte (5 829 923), todas na região Sudeste.^[3] Quase todas as capitais são as maiores cidades de seus estados, com exceção de Vitória, capital do Espírito Santo, e Florianópolis, a capital de Santa Catarina.^[174] Existem também regiões metropolitanas não capitais, como as de Campinas, Baixada Santista e Vale do Paraíba (todas no estado de São Paulo); Vale do Aço

(Minas Gerais); Serra Gaúcha (Rio Grande do Sul) e Vale do Itajaí (Santa Catarina).^[3]

Conurbações mais populosas do Brasil								
Estimativa populacional do Instituto Brasileiro de Geografia e Estatística (IBGE) para 1º de julho de 2015 ^{[175][176][177][nota 5]}								
	Posição	Localidade	Unidade federativa	Pop.	Posição	Localidade	Unidade federativa	Pop.
	1	São Paulo	São Paulo	21 015 117	11	Belém	Pará	2 126 518
	2	Rio de Janeiro	Rio de Janeiro	12 289 936	12	Campinas	São Paulo	2 058 598
	3	Belo Horizonte	Minas Gerais	5 069 897	13	Manaus	Amazonas	2 057 711
	4	Recife	Pernambuco	3 060 345	14	Vitória	Espírito Santo	1 700 200
								

Pesquisa em Dados na Web

- Tornar a Web um imenso BD! (Utopia?)
 - esquematização dos dados
 - consultas declarativas (estilo SQL, p.ex.)
- Para se alcançar este difícil objetivo...
 - descobrir onde estão os dados de interesse
 - extrair os dados de interesse
 - catalogar (esquematizar e persistir) e/ou indexar
- Tecnologias para se alcançar esse objetivo
 - dicionários, ontologias, bases de conhecimento, *machine learning*, reconhecimento de padrões, gramáticas, ...

Dados Semiestruturados (SEs)

- Foco de muitas pesquisas na área de gerenciamento de dados na Web
 - grande parte das “entidades” na Web tem natureza semiestruturada e está descrita em uma página ou em parte dela
- “Padrão” *de facto* para dados SEs: XML
 - formato capaz de representar dados SEs extraídos da Web
 - dados com representação heterogênea
 - dados com representação autodescritiva
 - dados com estrutura parcial

XML (*eXtensible Markup Language*)

- XML é uma metalinguagem de marcação
 - linguagem de marcação
 - semelhante à linguagem HTML
 - utiliza *tags* para descrição os dados
 - *tag*: indica a intenção do dado e delimita o seu conteúdo
 - metalinguagem
 - XML é um padrão aberto
 - cada aplicação define o protocolo (linguagem) para a representação dos seus dados de acordo com o significado (semântica) desejado para o dado

Estrutura Heterogênea

- Cada instância de dado pode ter um esquema particular

```
<autor>  
  <nome>Joao Silva</nome>  
  <endereco>rua B,23</endereco>  
  <eMail>jsilva@inf.ufsc.br</eMail>  
</autor>
```

```
<autor>  
  <nome>Ana Ramos</nome>  
  <endereco>  
    <rua>Brasil</rua>  
    <numero>767</numero>  
    <cidade>Fpolis</cidade>  
  </endereco>  
  <fone>33313333</fone>  
  <fone>33313332</fone>  
</autor>
```

Estrutura Autodescritiva

- Cada instância de dado carrega o seu esquema

```
<autor>  
  <nome>Ana Ramos</nome>  
  <endereco>  
    <rua>Brasil</rua>  
    <numero>767</numero>  
    <cidade>Fpolis</cidade>  
  </endereco>  
  <fone>33313333</fone>  
  <fone>33313332</fone>  
</autor>
```

Estrutura Parcial

- Apenas parte do conteúdo textual de uma instância pode apresentar uma estrutura

```
<capítulo numero = 2 titulo = "Tecnologia XML">  
  Este capítulo descreve ...  
  XML<ref>(Mel03)</ref>. XML é um padrão ...  
    <secao numero = 1>  
      <titulo>DTD</titulo>  
      Esta seção descreve ...  
    </secao>  
    ...  
</capítulo>
```

Dado XML: Dado Não-Convencional

```
<livro>  
  <titulo>Tecnologia XML</titulo> ← tag (intenção do dado)  
  <autor>  
    <nome>João da Silva</nome> ← conteúdo do dado  
    <eMail>js@hotmail.com</eMail>  
    <endereco>  
      <comercial>rua A, 34 - Fpolis - SC</comercial>  
      <residencial>rua B, 5 - Fpolis - SC</residencial>  
    </endereco>  
  </autor>  
  ...  
  <capitulo nome="Introdução">Este capítulo apresenta ...  
    <secao>  
      <nome>Linguagens de Marcação</nome>  
      ...  
    </secao>  
  </capitulo>  
  ...  
</livro>
```

→ estrutura hierárquica, ordenada e complexa

Sintaxe XML – Documento XML

- Dados XML são mantidos em um documento XML (.xml)
- Um documento XML geralmente contém
 - cabeçalho
 - dados (elementos, atributos e entidades)
 - comentários

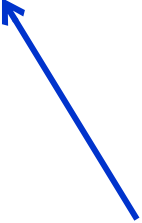
Exemplo de Documento XML

```
<?xml version ="1.0" encoding ="ISO-8859-1" ?>
<!-- documento XML sobre livros -->
<listaLivros>
<livro ISBN="112">
    <título>Tecnologia XML</título>
    <autor>
        <nome>João da Silva</nome>
        <eMail>js@hotmail.com</eMail>
    </autor>
    ...
    <capítulo nome="Introdução">A XML foi ...
        <seção>
            <nome>O uso do elemento <![CDATA[<?xml>]]></nome> ...
        </seção>
    </capítulo> ... <figura arquivo="exemplo.jpg"/>
</livro> ...
</listaLivros>
```

Exemplo de Documento XML

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!-- documento XML sobre livros -->
<listaLivros>
  <livro ISBN="112">
    <título>Tecnologia XML</título>
    <autor>
      <nome>João da Silva</nome>
      <eMail>js@hotmail.com</eMail>
    </autor>
    ...
    <capítulo nome="Introdução">A XML foi ...
      <seção>
        <nome>O uso do elemento <![CDATA[<?xml>]]></nome> ...
      </seção>
    </capítulo> ... <figura arquivo="exemplo.jpg"/>
  </livro> ...
</listaLivros>
```

Cabeçalho
(instrução de processamento –
inicia com <? e termina com ?>)



Exemplo de Documento XML

```
<?xml version ="1.0" encoding ="ISO-8859-1" ?>
<!-- documento XML sobre livros -->      ← comentário
<listaLivros>
  <livro ISBN="112">
    <título>Tecnologia XML</título>
    <autor>
      <nome>João da Silva</nome>
      <eMail>js@hotmail.com</eMail>
    </autor>
    ...
    <capítulo nome="Introdução">A XML foi ...
      <seção>
        <nome>O uso do elemento <![CDATA[<?xml>]]></nome> ...
      </seção>
    </capítulo> ... <figura arquivo="exemplo.jpg"/>
  </livro> ...
</listaLivros>
```

Exemplo de Documento XML

```
<?xml version ="1.0" encoding ="ISO-8859-1" ?>
<!-- documento XML sobre livros -->
<listaLivros>
  <livro ISBN="112">
    <título>Tecnologia XML</título>
    <autor>
      <nome>João da Silva</nome>
      <eMail>js@hotmail.com</eMail>
    </autor>
    ...
    <capítulo nome="Introdução">A XML foi ...
      <seção>
        <nome>O uso do elemento <![CDATA[<?xml>]]></nome> ...
      </seção> <figura arquivo="exemplo.jpg"></>
    </capítulo> ...
  </livro> ...
</listaLivros>
```

elemento raiz

elemento simples

elemento complexo

elemento misto

elemento vazio

Exemplo de Documento XML

```
<?xml version ="1.0" encoding ="ISO-8859-1" ?>
<!-- documento XML sobre livros -->
<listaLivros>
  <livro ISBN="112">
    <título>Tecnologia XML</título>
    <autor>
      <nome>João da Silva</nome>
      <eMail>js@hotmail.com</eMail>
    </autor>
    ...
    <capítulo nome="Introdução">A XML foi ...
      <seção>
        <nome>O uso do elemento <![CDATA[<?xml>]]></nome> ...
      </seção>
    </capítulo> ...
  </livro> ...
</listaLivros>
```

← atributo

Tecnologias para XML

- Recursos para gerenciamento de dados XML
 - regulamentados pelo consórcio W3C
 - *W3C (World Wide Web Consortium)* – <http://www.w3.org/>
- Principais tecnologias
 - definição de esquemas
 - DTD e XML Schema
 - linguagens de consulta
 - XPath e XQuery
 - modelo de representação e API de acesso
 - DOM

Definição de Esquemas

- Esquema XML

- define restrições para a **organização hierárquica** e **conteúdo** dos dados em um doc XML
- **documento válido**
 - documento cuja estrutura está de acordo com um esquema
 - validação é feita por um *parser*

- Duas recomendações

- **DTD** (*Document Type Definition*)
- **XSD** (*XML Schema Definition*)

DTD

- Primeira recomendação da W3C
- **Gramática** para definição de hierarquia
 - baseada em seqüências ordenadas e escolhas
- **Definição de elementos**
 - complexos, textuais (`#PCDATA`), vazios (`EMPTY`), mistos (`((#PCDATA | ...) *)`) ou com conteúdo aberto (`ANY`)
- **Definição de atributos**
 - obrigatórios (`#REQUIRED`) opcionais (`#IMPLIED`), fixos (`#FIXED`), valor *default*, enumeração, referência (`ID`, `IDREF(S)`)

DTD - Exemplo

```
<!ELEMENT listaLivros (livro+)>
<!ELEMENT livro (título, preço, autor+,
                 capítulo+)>
<!ATTLIST livro ISBN ID #REQUIRED
               edicaoAnterior IDREF #IMPLIED>
<!ELEMENT título (#PCDATA)>
<!ELEMENT autor (nome, eMail?)>
<!ELEMENT nome (#PCDATA)>
<!ELEMENT preço (#PCDATA)>
<!ELEMENT eMail (#PCDATA)>
<!ELEMENT capítulo (#PCDATA | seção)*>
<!ATTLIST capítulo nome CDATA #REQUIRED>
<!ELEMENT seção (nome, conteúdo)>
<!ELEMENT conteúdo (#PCDATA)>
```

XML Schema (XSD)

- Recomendação mais recente
- Sintaxe XML
- Extensão da funcionalidade de uma DTD
 - definição e especialização de tipos de elementos
 - semelhança com esquemas orientados a objetos
 - definição de tipos de dados
 - simples (*string, integer, boolean, ...*)
 - complexos (*list, union*)
 - facilidades adicionais para definição de restrições
 - intervalos de valores permitidos, padrões de conteúdo via expressões regulares, ...
 - ...

XSD - Exemplo

```
<?xml version="1.0" encoding="UTF-8">
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  ... <!-- Declaração de Tipos -->
  <xsd:simpleType name="Tisbn">
    <xsd:restriction base="xsd:string">
      <xsd:pattern value="[0-9]{2}-[0-9]{3}-[0-9]{4}-[0-9]"/>
    </xsd:restriction>
  </xsd:simpleType>
  <xsd:complexType name="Tlivro">
    <xsd:sequence>
      <xsd:element name="titulo" type="xsd:string"/>
      <xsd:element name="autor" type="Tautor"
        minOccurs="1" maxOccurs="unbounded"/>
      <xsd:element name="preço" type="xsd:float"/>
      ...
    </xsd:sequence>
    <xsd:attribute name="isbn" type="Tisbn"/>
  </xsd:complexType>
  ...
```

XSD – Exemplo (cont.)

...

```
<xsd:complexType name="TlivroTécnico" base="Tlivro"
    derivedBy="extension">
  <xsd:element name="area" type="xsd:string"
    minOccurs="1" maxOccurs="1"/>
</complexType>
...
<!-- Declaração de Elementos -->
<xsd:element name="listaLivros">
  <xsd:complexType>
    <xsd:sequence>
      <xsd:element name="livro" type="Tlivro"
        minOccurs="1" maxOccurs="unbounded"/>
    </xsd:sequence>
  </xsd:complexType>
</xsd:element>
</xsd:schema>
```

XPath

- Primeira recomendação para consulta a dados
- Linguagem para o acesso a partes de um doc XML
 - sintaxe: expressões de caminho
 - assemelha-se à navegação em diretórios de arquivos
 - exemplo
 - expressão XPath: `/listaLivros/livro/título`
 - resultado:
`<título>Tecnologia XML</título>`
`<título>Sistema de Banco de Dados</título>`
`...`

XPath - Exemplos

`/listaLivros`

`/listaLivros/livro/*/eMail`

`/listaLivros/livro//seção`

`/listaLivros/livro/capítulo[1]`

`/listaLivros/livro/capítulo/nome |`

`/listaLivros/livro/capítulo/seção/nome`

`/listaLivros/livro/@ISBN`

`/listaLivros/livro[título = "XML"]`

`/listaLivros/livro[capitulo/@nome = "XML" or //seção/nome
= "XML"]/título`

`/listaLivros/livro//seção[last()]`

XQuery

- Recomendação mais recente
- Recursos adicionais em relação à *XPath*
 - junções, definição de estruturas de resultado, variáveis de consulta, atributos calculados, funções de agregação, ...
- Sintaxe básica

```
for variável in expressãoXPath  
[let associação de novas variáveis]  
[where condição]  
return estrutura de resultado
```

XQuery - Exemplos

```
for $liv in /listaLivros/livro
where $liv/autor/nome = "João Silva"
return { $liv/@ISBN, $liv/titulo }
```

(consulta
simples)

```
for $liv in /listaLivros/livro
let $pDesc := $liv/preço - $liv/preço * 0.1
where $liv/categoria = "ficcao"
return <FiccaoDesc>{$liv/titulo, $pDesc}</FiccaoDesc>
```

```
for $liv1 in /listaLivros/livro[@ISBN = "562"]
for $liv2 in /listaLivros/livro
where $liv2/@ISBN != $liv1/@ISBN
and $liv2/autor/nome = $liv1/autor/nome
return $liv2/titulo
```

(nova
estrutura
de
resultado)

(junção)

DOM (*Document Object Model*)

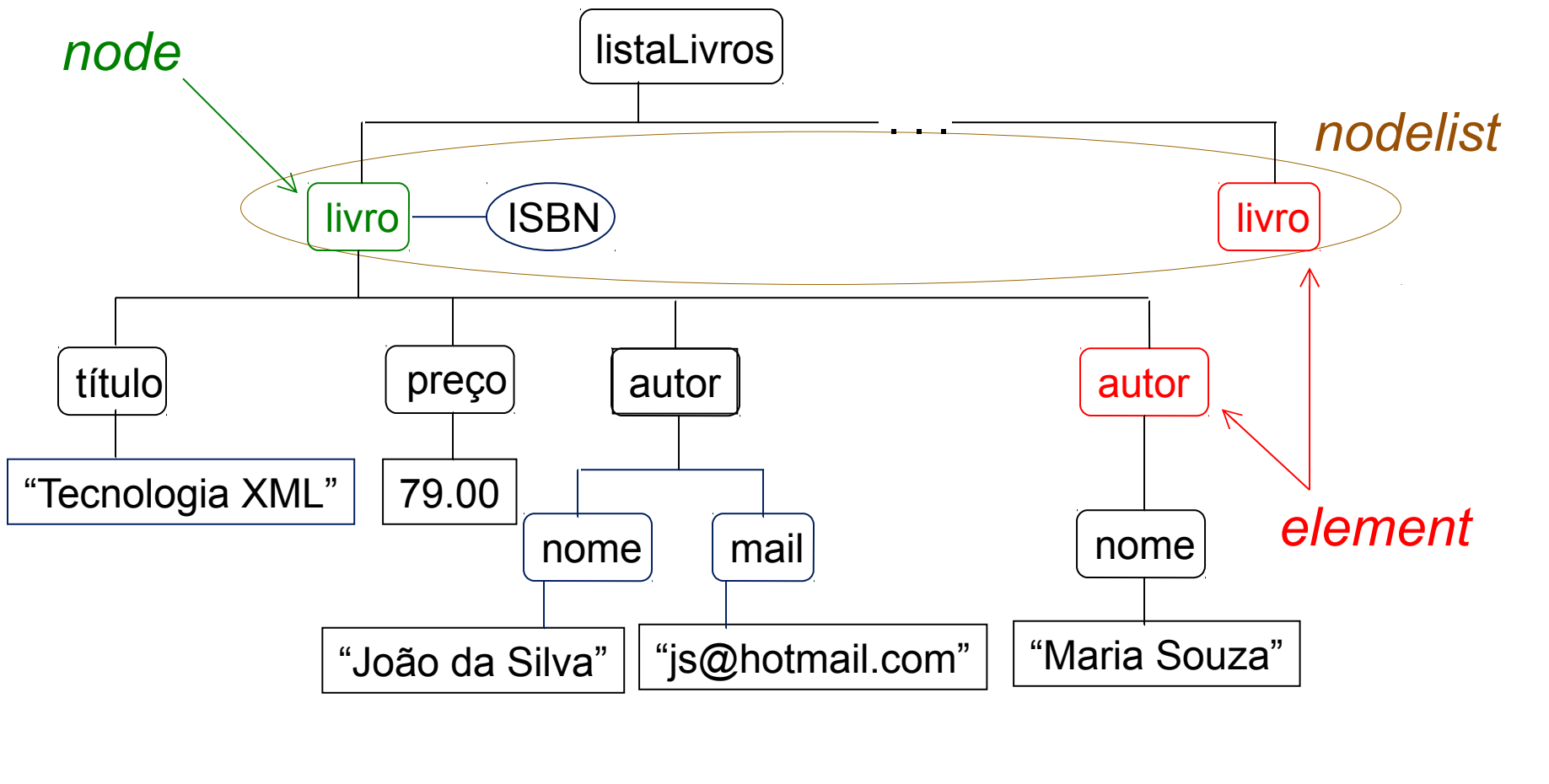
- Modelo de representação para dados XML (para fins de processamento)
 - estrutura hierárquica (árvore)
 - métodos de acesso (API DOM)
 - principais classes de objetos
 - *document*, *node*, *nodelist* e *element*
 - execução de consultas e atualizações de dados
- *Parsers* DOM
 - validam um doc XML
 - geram um objeto *document*

Objetos do Modelo DOM

document

node

nodelist



Exemplos de Métodos - *Node*

Método	Resultado
<i>nodeName</i>	string
<i>nodeValue</i>	string
<i>nodeType</i>	short
<i>parentNode</i>	Node
<i>childNodes</i>	NodeList
<i>firstChild</i>	Node
<i>lastChild</i>	Node
<i>previousSibling</i>	Node
<i>nextSibling</i>	Node
<i>insertBefore(Node novo, Node ref)</i>	Node
<i>replaceChild(Node novo, Node antigo)</i>	Node
<i>removeChild(Node)</i>	Node
<i>hasChildNode</i>	boolean

Exemplos de Métodos - *Element*

Método	Resultado
<i>tagName</i>	string
<i>getAttribute(String)</i>	string
<i>setAttribute(String nome, String valor)</i>	Attr
<i>getAttributeNode(String)</i>	Attr
<i>removeAttributeNode(String)</i>	Attr
<i>getElementsByTagName</i>	NodeList

Exemplos de Métodos - *Nodelist*

Método	Resultado
<i>Length</i>	int
<i>item(int)</i>	Node

DOM – Exemplo (*JavaScript*)

```
var doc, raiz, livro1, autores, autor2;
doc = new ActiveXObject("Microsoft.XMLDOM");
doc.load("livros.xml");
if (doc.parseError != 0) ...;
else
{
    raiz = doc.documentElement;
    /* busca o primeiro livro (primeiro nodo filho) */
    livro1 = raiz.childNodes.item(0);
    /* busca a lista de autores do primeiro livro */
    autores = livro1.getElementsByTagName("autor");
    /* busca o segundo autor */
    autor2 = autores.item(1);
    /* escreve o nome do autor - primeiro nodo filho */
    document.write("Nome do segundo autor: " +
        autor.childNodes.item(0).data);
}
```