Arquivos XML

Prof. Walmes Zeviani

walmes@ufpr.br

Laboratório de Estatística e Geoinformação
Departamento de Estatística
Universidade Federal do Paraná

Introdução

Motivação

1. XML é uma esquema geral para representar de dados.
2. Vários tipos de arquivos são variações de XML.
3. É amplamente usado em Web APIs.
4. Possui linguagem de consulta própria.
5. É um padrão da W3C.

Objetivos

1. Fazer a definição e discutir os principais aspectos do XML.
2. Apresentar alguns dialetos XML.
3. Dar uma visão geral da estrutura e sintaxe do XML.

XML

Definição e características

- XML: *eXtensible Markup Language*.
- Usado para representar dados em diversos formatos.
- Exemplo: tabelas, planilhas, documentos de texto, imagens, estilos de citação, etc.
- É tão genérico que pode representar qualquer tipo de dado.

Dialetos de XML

- HTML (*HiperText Markup Language*): páginas de internet.
- KML (*Keyhole Markup Language*): informação geográfica tri-dimensional.
- CSL (*Citation Style Language*): referências bibliográficas.
- ODF (*Open Document Format*): documentos de texto, planilha e slides, etc.
- SVG (*Scalable Vector Graphics*): formato de imagens vetoriais.
- Epub: publicação/livro eletrônico.

```xml
<?xml version="1.0"?>
<catalog>
   <book id="bk101">
      <author>Gambardella, Matthew</author>
      <title>XML Developer's Guide</title>
      <genre>Computer</genre>
      <price>44.95</price>
      <publish_date>2000-10-01</publish_date>
      <description>An in-depth look at creating applications
      with XML.</description>
   </book>
   <book id="bk102">
      <author>Ralls, Kim</author>
      <title>Midnight Rain</title>
      <genre>Fantasy</genre>
      <price>5.95</price>
      <publish_date>2000-12-16</publish_date>
      <description>A former architect battles corporate zombies,
      an evil sorceress, and her own childhood to become queen
      of the world.</description>
   </book>
   <book id="bk103"> ...
   <book id="bk112"> ...
   </book>
</catalog>
```

Figura 1. Fragmento de arquivo XML que contém um catálogo de livros. Fonte: <https://msdn.microsoft.com/en-us/library/ms762271(v=vs.85).aspx>.

```
<!DOCTYPE kml>
<kml xmlns="">
  <Document>
    <Placemark>
      <LineString>
        <coordinates>-59.9729450,-11.9763182 -59.9729413,-11.9763182 -59.969472
      </LineString>
    </Placemark>
    <Placemark>
      <LineString>
        <coordinates>-46.8062022,-15.8706177 -46.8050041,-15.8657058 -46.805501
      </LineString>
    </Placemark>
    <Placemark>
      <LineString>
        <coordinates>-49.3352376,-24.2290424 -49.3347402,-24.2310370 -49.330192
      </LineString>
    </Placemark>
```

Figura 2. Fragmento de arquivo KML que contém as delimitações geográficas dos Estados brasileiros. Fonte: <http://www.gmapas.com/poligonos-ibge/poligonos-estados-do-brasil>.

```
<?xml version="1.0" encoding="utf-8"?>
<style xmlns="http://purl.org/net/xbiblio/csl" class="in-text" version="1.0">
  <info> ...
  </info>
  <macro name="container-contributors"> ...
  </macro>
  <macro name="secondary-contributors"> ...
  </macro>
  <macro name="author">
    <names variable="author">
      <name name-as-sort-order="all" sort-separator=", " initialize-with=". " delimiter="; " delimiter-precedes-last="always">
        <name-part name="family" text-case="uppercase"/>
        <name-part name="given" text-case="uppercase"/>
      </name>
      <label form="short" prefix=" (" suffix=".)" text-case="uppercase" strip-periods="true"/>
      <substitute>
        <names variable="editor"/>
        <names variable="translator"/>
        <text macro="title"/>
      </substitute>
    </names>
  </macro>
  <macro name="author-short">
```

Figura 3. Fragmento de arquivo CSL que contém as instruções para formatação de referências bibliográficas nas normas da ABNT. Fonte: <https://metodologiaetecnologia.com.br/2011/09/26/mendeley-parte-12-estilos-abnt-vancouver-e-outros/>.

Autópsia do XML

Arquivos XML e dialetos online

Exemplos online

- Catálogo botânico: https://www.w3schools.com/Xml/plant_catalog.xml.
- Catálogo de livros: https://msdn.microsoft.com/en-us/library/ms762271(v=vs.85).aspx.
- Polígonos dos Estados: http://www.gmapas.com/poligonos-ibge.
- Imagem vetorial: https://upload.wikimedia.org/wikipedia/en/2/22/Heckert_GNU_white.svg.

Embelezadores de XML

- https://codebeautify.org/xmlviewer.
- https://jsonformatter.org/xml-viewer.
- https://countwordsfree.com/xmlviewer.

Essencial de XML

- A unidade básica é o **elemento** ou **nó** (*node*).
- O elemento é começa e termina com a **tag nomeada**.
  - `<title>Os Três Mosqueteiros</title>`.
  - `<BOTANICAL>Eschscholzia californica</BOTANICAL>`.
- O **par de tags** delimita o conteúdo do elemento.
- O conteúdo de um elemento pode ser outros elementos.
- Um elemento sempre tem nome.
- Um elemento pode ter **atributos**.
- Atributos são do tipo `campo = "valor"`.

Estrutura do XML

- Elementos podem conter elementos → **estrutura hierárquica**.
- Atributos registram **metadados** ou informações sobre os dados. Ex: estampa de tempo, classe, id.
- Capaz de representar **estruturas complexas** de dados.
- A estrutura se assemelha a de árvore.
- Embora muito verboso, a **taxa de compressão** é boa.

Regras

- Sempre ter um elemento raíz, declaração e instruções de processamento.
- Os elementos devem estar apropriadamente aninhados.
- As tags são *case sensitive*.
- Tags devem ocorrer aos pares ou ser auto contida.
  - `<img src="image.png"></img>`
  - `<img src="image.png"/>`
- Atributos são sempre na *tag* de abertura.
- Não usar espaços nas estremidades das tags.
- Atributos são no formato `campo = "valor"`.
  - O valor deve estar entre aspas.
  - Nome do atributo devem ser único no elemento.
- Nomes dos elementos podem iniciar com `[_A-Za-z]`.
- O sinal de `:` é reservado para *namespaces*.
- A palavra `xml` é reservada, não pode ser usada.
- Os caracteres `&`, `<` e `>` como valor devem ser codificados.

Pontos contra o XML

- Por causa das tags nomeadas, o XML muito verboso.
- XML é texto pleno e *human-readable*.
- Ferramentas de visualização e edição são muito desejáveis.
- Veja os embelezadores online de XML.
- Editores de texto geral conseguem “embelezar” arquivos XML.

Estrutura hierárquica

Modelo de árvore

- O modelo de **árvore** é útil para processamento e navegação.
- Um dos mecanismos de parsing é o **DOM** (*Document Object Model*).
- O DOM analisa (*parsing*) o XML e cria uma estrutura em forma de árvore que representa o documento.
- Com a estrutura criada é possível caminhar e extrair conteúdo de forma simples.
- A linguagem **XPath** é usada para fazer consulta no DOM.

Terminologia

- Elementos são nós da árvore.
- Conexões são ramos da árvore.
- Relação entre nós expressa por relações famíliares.
- *root*: nó raíz, o mais alto e origem da árvore.
- *branch*: ramo que conecta dois nós.
- *parent/child*: relação de parentesco entre nós onde *parent* é o que está mais próximo da origem.
- *ancestor/decendent*: relações de parentesco verticais entre nós separados por outros nós.
- *sibling*: relação horizontal, nós que possuem o mesmo *parent*.
- *leaf nodes* ou terminais: nós que não possuem *child*.
- Profundidade é a distância com relação ao *root* em número de conexões.

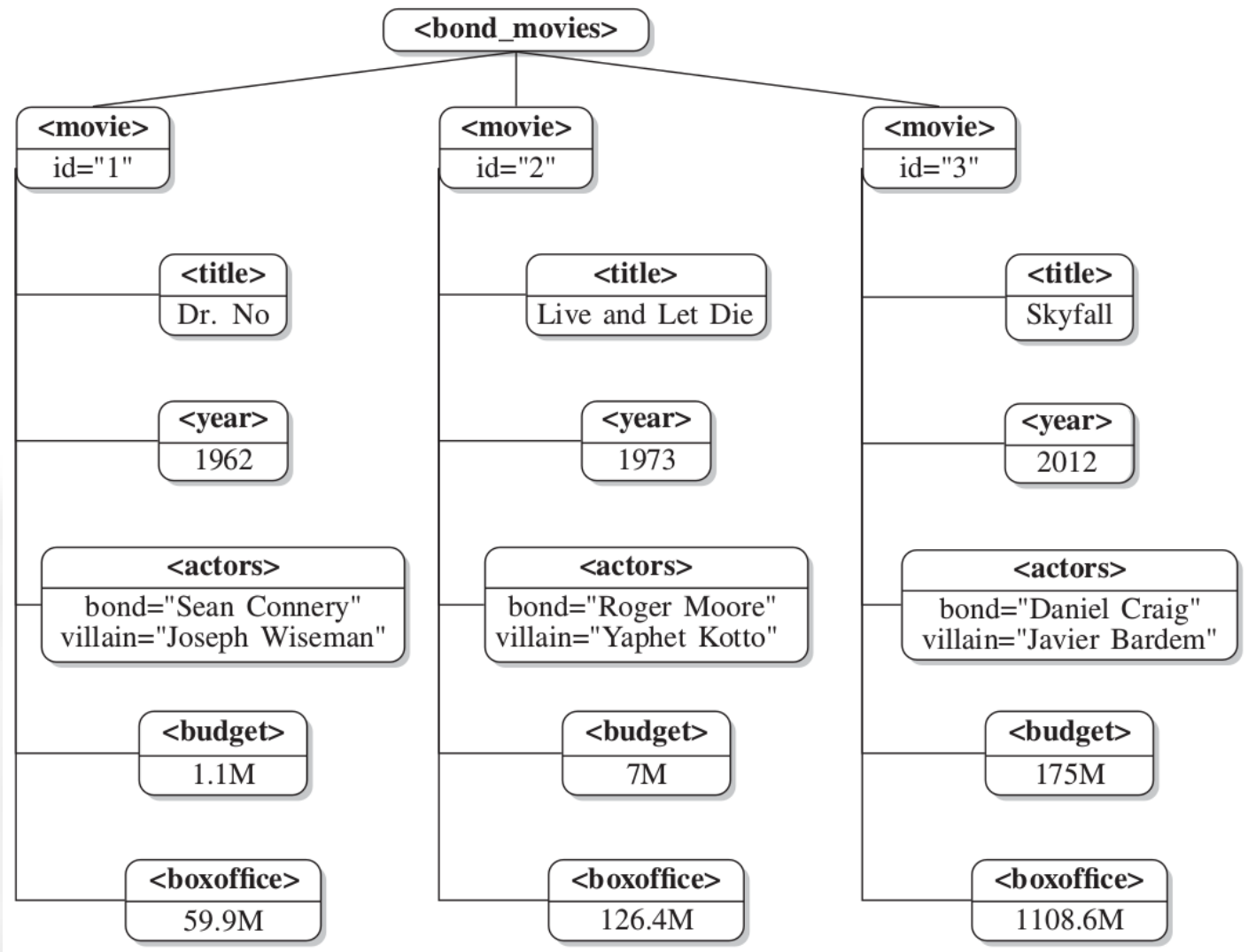

Figura 4. Exemplo de arquivo XML em forma de árvore. Fonte: MUNZERT et al. (2015), página 47.

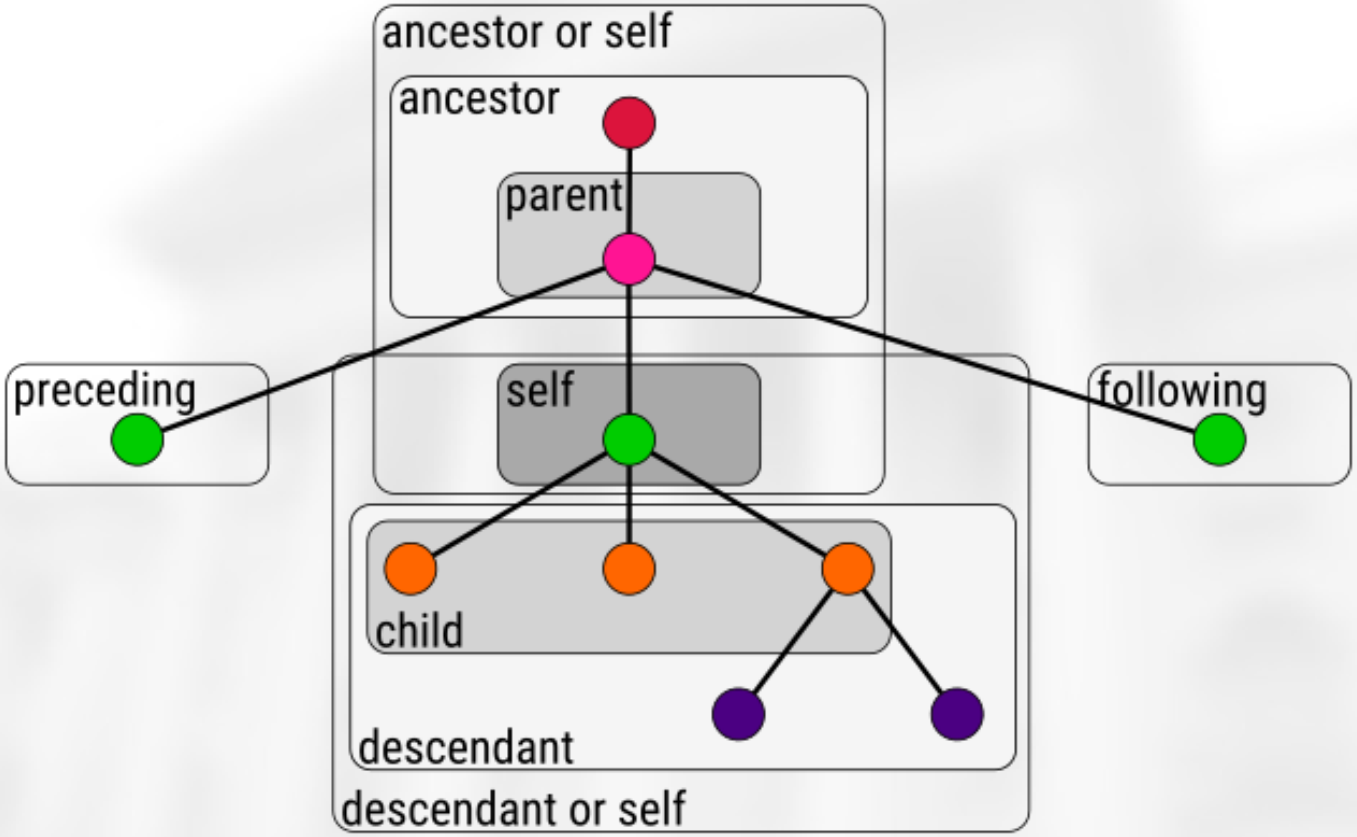

Figura 5. Relações familiares entre nós de um arquivo XML com a representação DOM.
Fonte: MUNZERT et al. (2015), página 87.

Resumo

- XML é um padrão de representação com estrutura geral para descrever conteúdo diverso.
- O XML emprega um esquema para especificar elementos, atributos e conteúdo em organização hierárquica.
- É capaz de representar diversos tipos de dados.
- Existem vários dialetos XML.
- Um arquivo XML tem estrutura de árvore.
- A posição relativa entre os nós emprega terminologia de árvore genealógica.

Referências

MUNZERT, S.; RUBBA, C.; MEIßNER, P.; NYHUIS, D. **Automated data collection with R: A practical guide to web scraping and text mining**. Wiley, 2015.