Integração de Sistemas e XML

Prof. Sergio Serra

sergioserra@gmail.com

Introdução ao XML

Aula 01

Universidade Federal
do Rio de Janeiro

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

    CREATE TABLE empregado (matricula int, nome varchar(30),
    salario float, depto int)

- Dados não estruturados
  - Não há nenhuma estrutura prévia
    Ex.: imagem, video, áudio, etc.

Dados Semi-estruturados

Características

- 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}

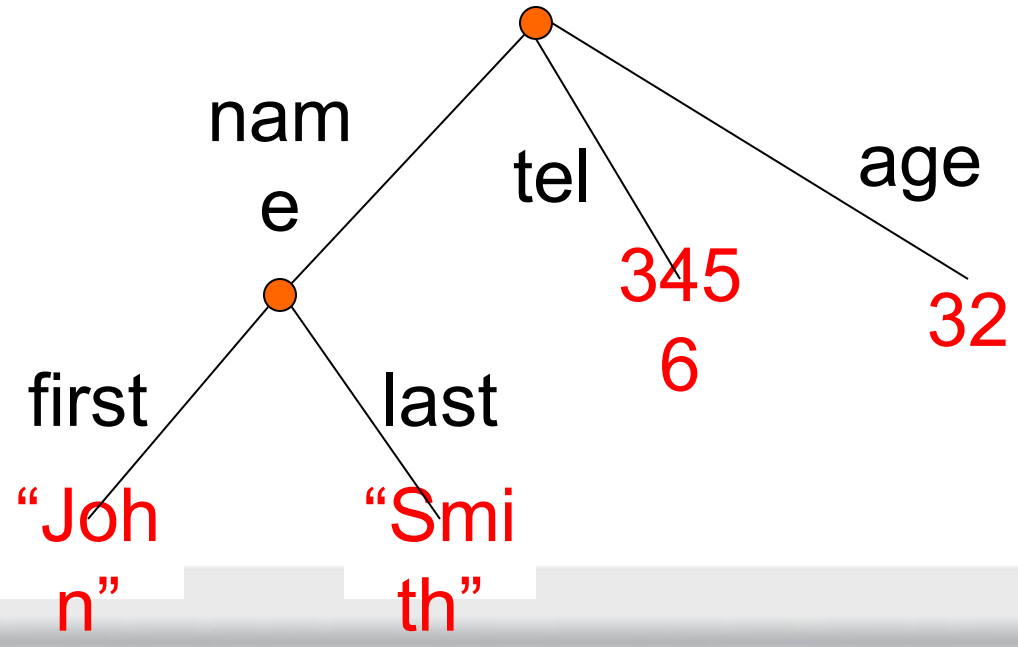
– rótulos de atributo não necessariamente únicos

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

Dados Semi-estruturados

- Podem ser representados graficamente
  - nós representam objetos conectados por arestas que os descrevem

Ex.: {name: {first:“John”, last: “Smith”}, tel: 3456, age: 32}



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...

Exemplos em SI

- Arquivos BibTex
  - Têm estrutura mas não é regular
  - Alguns atributos não aparecem, apesar de obrigatórios

```
@article{Gettys90,
    author = {Jim Gettys and Phil
    Karlton and Scott McGregor},
    title = {The X Window System,
    Version 11},
    journal = {Software Practice and
    Experience},
    volume = {20},
    number = {S2},
    year = {1990},
    postscript =
    “papers/gettys90.ps.gz”,
    abstract = {A technical overview of
    the X11 functionality. This is an
    update of the X10 TOG paper by
    Scheifler \& Gettys.}
}
```

Exemplos em SI

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  MEDLINE   95176709
  PUBMED    7871890
FEATURES             Location/Qualifiers
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
```

Arquivos GenBank

Exemplos

- Guia de restaurantes (Palo Alto Weekly newspaper)
  - Cada restaurante apresenta uma estrutura diferente

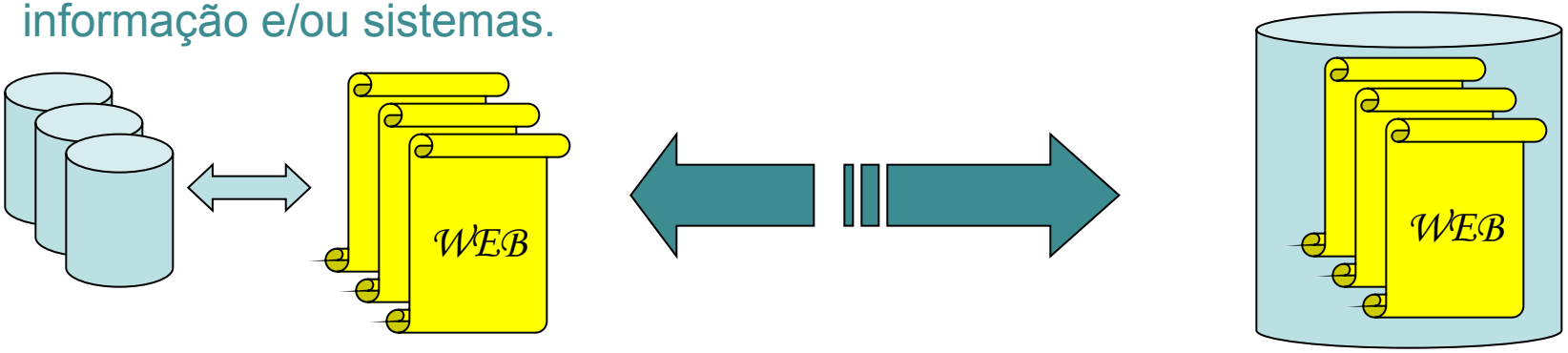
```
Guide
Restaurant
Name "Blues on the Bay"
Category "Vegeterian"
Entree
Name "Black bean soup"
Price "10.00"
Entree
Name "Asparagus Timbale"
Price "22.50"
Location
Street "1890 Wharf Ave"
City "San Francisco"
Restaurant
Name "McDonald's"
Category "Fast Food"
Price "cheap"
Nearby "Blues on the Bay"
```

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 uma grande base de dados.



Descrever os dados da Web

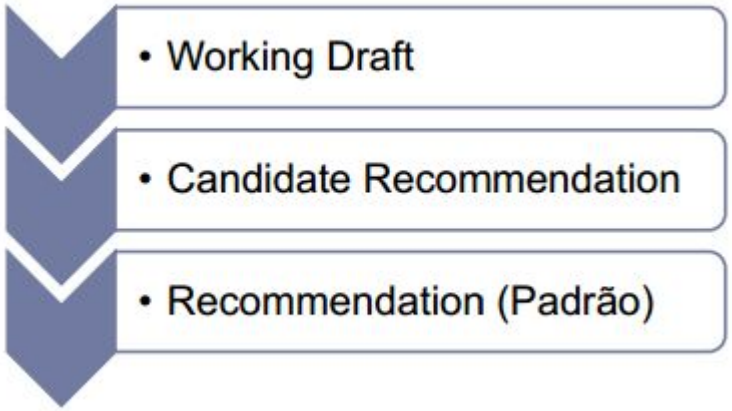
- Tratar dados semi-estruturados
- Separar o conteúdo:
  - Independência de armazenamento
    - Permite a visualização de dados provenientes de fontes heterogêneas
  - Independência de apresentação
    - Permite que as aplicações apresentem/tratem os dados como lhes é conveniente

O que é XML?

- e**X**tensible **M**arkup **L**anguage
- Linguagem padrão para marcação de dados na Web, com foco na descrição do conteúdo
- Idealizada pelo W3C (**www.w3c.org**)
- Influenciou um conjunto de tecnologias derivadas:
  - Xlink, Xpointer, Xschema, XSLT, DOM, SAX, XSL, XML Namespaces, Efficient XML Interchange (EXI)…

O que é W3C?

- Consórcio internacional responsável pela padronização de iniciativas ligadas à Web
  - Ex.: HTML, XML e iniciativas relacionadas, entre outros
- Especificações dessas iniciativas são classificadas de acordo com seu nível de "maturidade



XML ≠ HTML

- XML – descreve o **conteúdo** do documento
  - Usuário define suas próprias tags para criar uma estrutura
  - Um documento XML não tem nenhuma instrução para apresentação
- HTML – descreve o **formato** do documento
  - HTML tem um conjunto fixo de tags e não descreve conteúdo
  - Um documento HTML contém instruções de representação

Histórico do XML

- 1994: primeiros trabalhos sobre adaptação das técnicas **SGML** à Web.
  - *'HTML to the Max: A Manifesto for Adding SGML Intelligence to the World Wide Web'* (Sperberg-McQueen e Goldstein ,1994)
  - O *Standard Generalized Markup Language* (**SGML**) é uma metalinguagem através da qual se pode definir linguagens de marcação para documentos.
- Junho 1996: criação de um grupo de trabalho no W3C

Histórico do XML

- 1996
  - 80 peritos em SGML uniram forças ao W3C (World Wide Web Consortium)
  - Objetivo: Definir uma linguagem de marcação com o poder da SGML, porém fácil de ser implementada
  - Forte influência do LOREL (*Lightweight Object Repository Language*) (Abiteboul et al, 1996)
- 10 fevereiro 1998
  - publicação da recomendação para versão 1.0 da linguagem

SGML - Características

- "Standard Generalized Markup Language"
- Uma linguagem de marcação abrangente mas complexa
- Desenvolvida por Charles **G**oldfarg, Edward **M**osher e Raymond **L**orie
- Adequada para aplicações envolvendo documentos grandes e complexos
- Tornou-se um padrão ISO (ISO 8879) na década de 80
- XML usa 10% de SGML para representar de forma eficaz 90% dos documentos

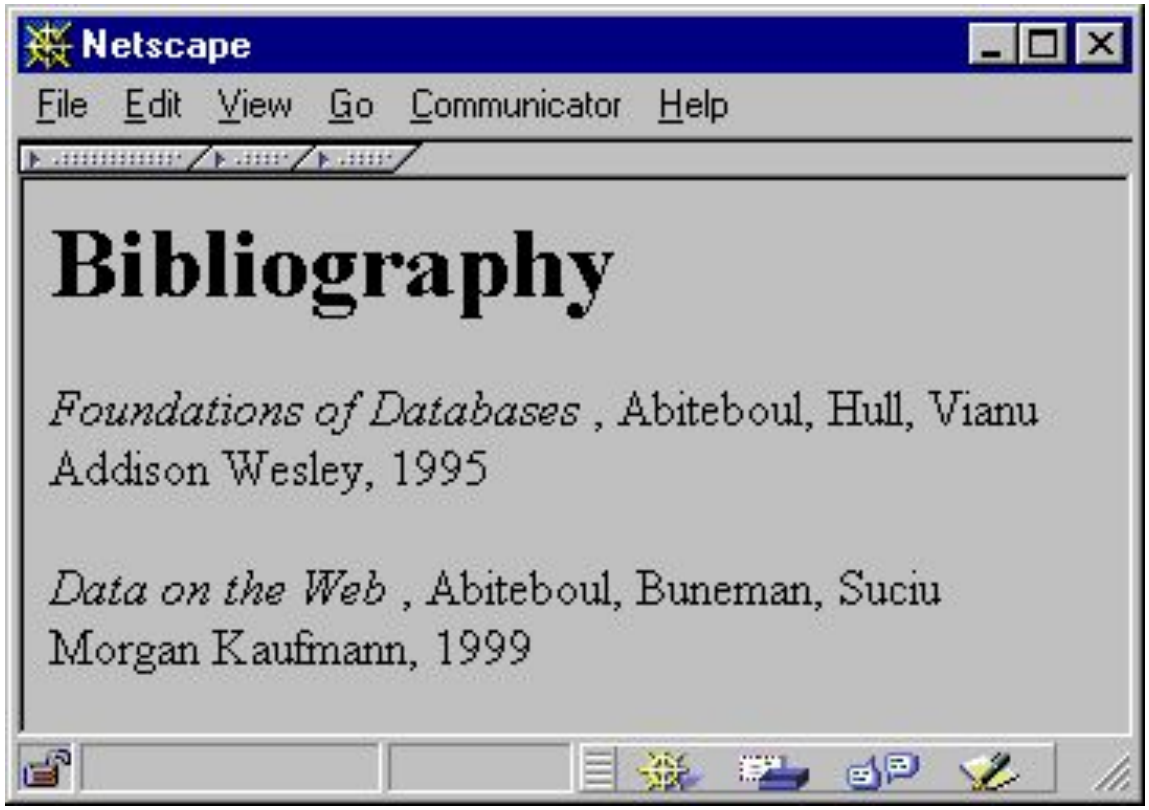
XML vs. HTML

- Usuário define suas próprias tags para criar uma estrutura
  - Mais prolixa que o HTML
- Estruturas podem ser aninhadas em um nível de profundidade arbitrário
- Um documento XML não tem nenhuma instrução para apresentação
- XML $\subset$ SGML mas HTML $\in$ SGML
- Um documento XML pode conter uma descrição opcional de sua estrutura (**DTD, XML Schema**)

Linguagens de Marcação

- SGML – linguagem de marcação com regras para definição de **classes** de documentos

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

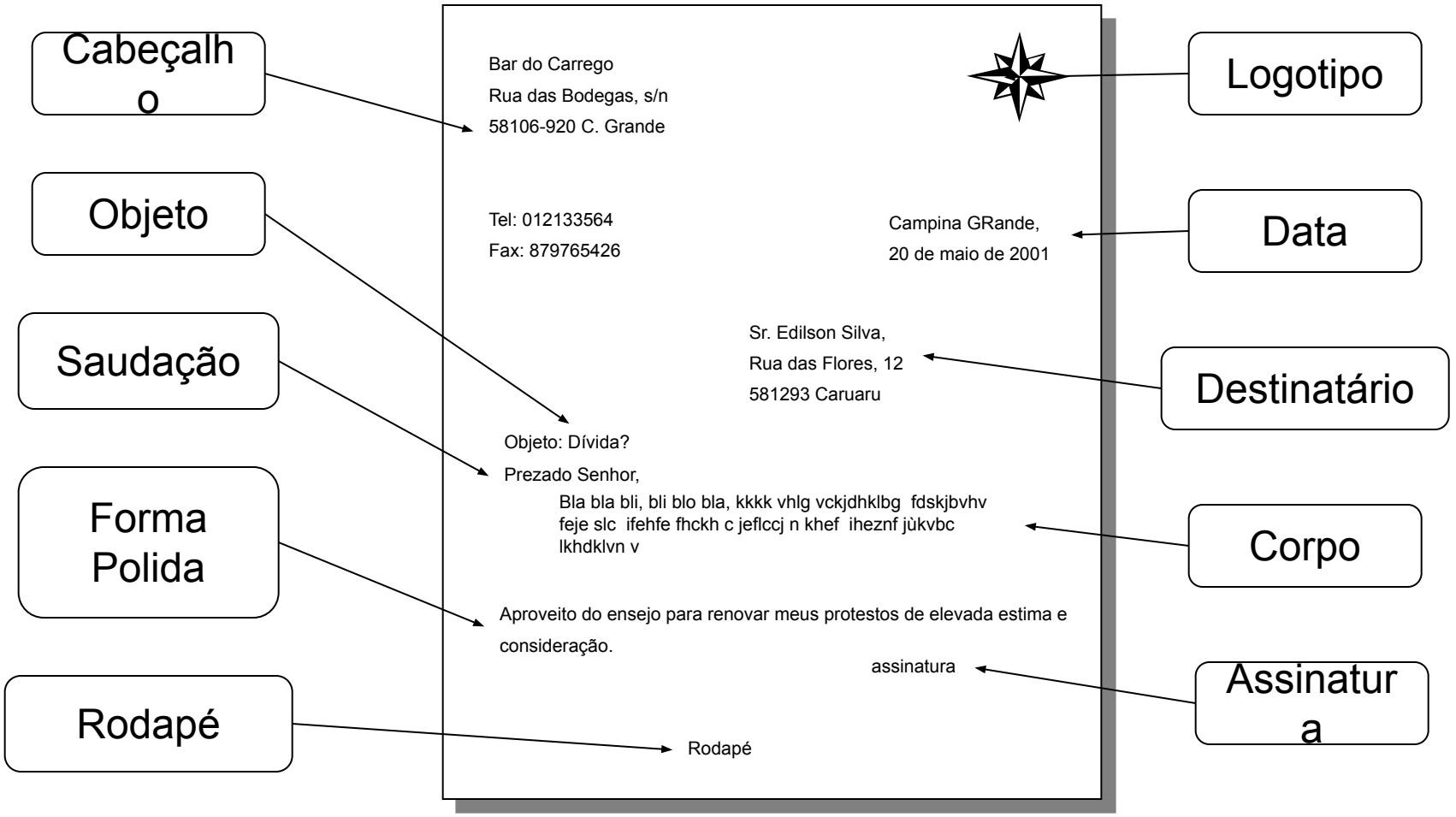
XML: dimensões e processamento ...

- 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


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

```
<cabecalho>
  . . .
</cabecalho>
<destinatario>
  <nome> Sr Edilson Silva </nome>
  <endereco>
    <rua> rua das Flores </rua>
    <cidade> Caruaru </cidade>
  </endereco>
</destinatario>
<objeto> bla bla </objeto>
<data> 20 Maio 2001 </data>
<saudacao> Prezado Senhor, </saudacao>
<corpo>
  <para>Aqui é o primeiro parágrafo</para>
  <para> aqui é o segundo ... </para>
</corpo>
```

```
<logotipo loc="logo-graph"/>
    <endereco> &abrev-endereco;
</endereco>
```

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?