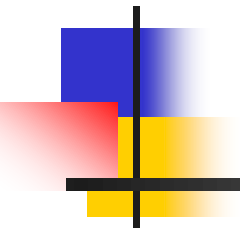
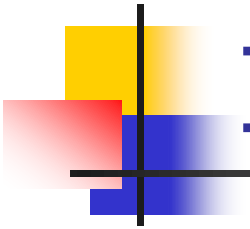


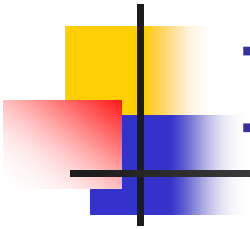
Natureza dos Dados

Adaptado de C. Batista



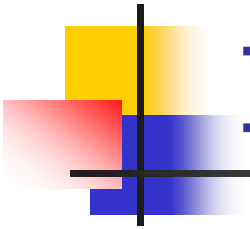
Introdução

- Revisando SGBD
  - Qual o propósito de um SGBD?
    - Independência de dados-programa
    - Persistência de dados
    - Concorrência
    - Recuperação de falhas
    - Processamento de consultas
    - Controle de integridade
    - Controle de Segurança
    - Distribuição dos dados



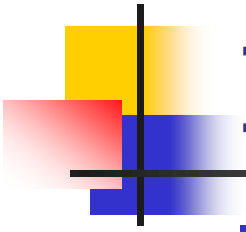
Introdução

- Revisando SGBD
  - Características:
    - Esquema pré-definido
    - Esquema pouco muda uma vez definido
    - Estrutura rígida
    - Níveis de abstração
    - Linguagem de alto nível declarativa para definição e manipulação dos dados (DDL e DML)



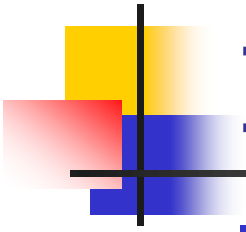
Introdução

- Dados
  - Estruturados
  - Semi-estruturados
  - Não-estruturados



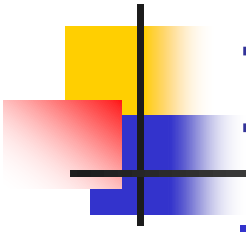
Introdução

- Dados Estruturados
  - São os dados armazenados no SGBD's com esquema rígido.
  - Ex.: Create table Empregado (matricula int, nome varchar(30), salario float);
  - A estrutura é conhecida a priori, então os comandos de inserção, seleção, atualização e remoção usam esta estrutura para manipular os dados
  - Neste caso a integração com a Web se dá através do uso de um protocolo de conectividade, por exemplo JDBC.



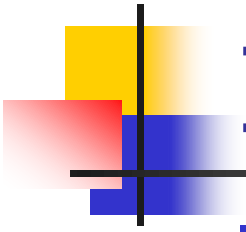
Introdução

- Dados Não-Estruturados
  - São dados que não possuem nenhuma estrutura, tais como um texto, uma imagem, um video. Portanto, do ponto de vista do SGBD estes dados são considerados como uma caixa preta.
  - (fluxo de bytes).



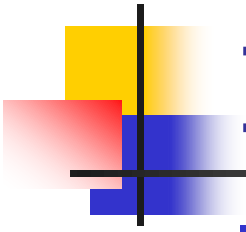
Introdução

- Dados Semi-Estruturados
  - São conhecidos pela ausência de esquema e por serem auto-descritivos.
  - Apresentam uma representação estrutural heterogênea



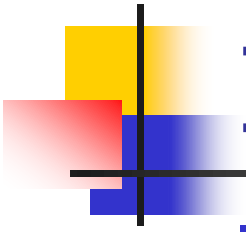
Introdução

- Dados Semi-Estruturados
  - Possuem as seguintes características:
    - Definição à posteriori
    - Estrutura irregular (ex. curriculum vitae)
    - Estrutura implícita
    - Estrutura parcial
    - Estrutura extensa
    - Estrutura evolucionária
    - Distinção entre estrutura e dados não é clara



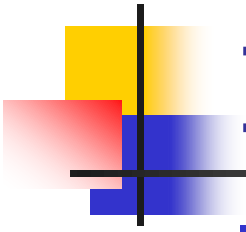
Introdução

- 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 possuem atributos a mais



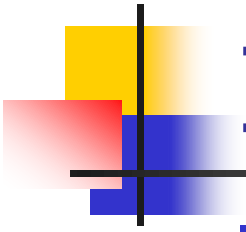
Introdução

- Dados Incompletos
  - Nem todo endereço tem caixa postal
  - Nem todo livro tem apêndice ou prefácio



Introdução

- Dados Incompletos
  - Sua estrutura não é previamente conhecida, pode não existir à parte
  - São auto-descritivos, i.e., embute a própria estrutura



Introdução

- Auto Descrição

  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

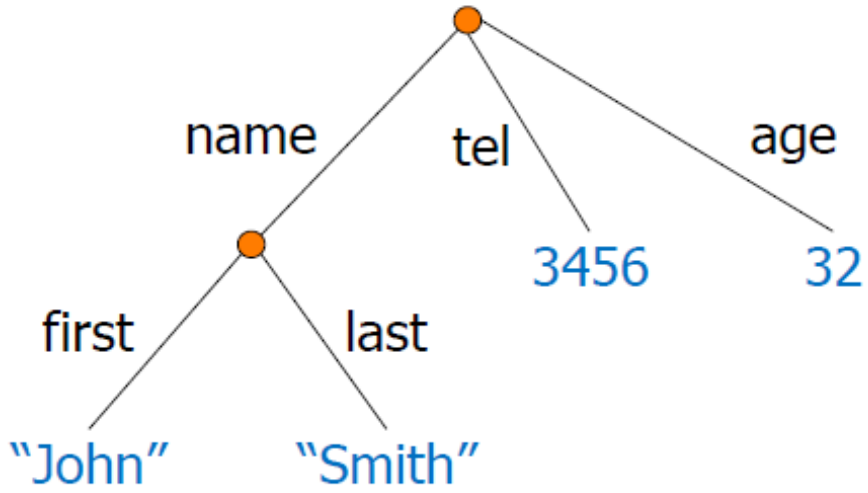
  {name: “John Smith”, tel: 3456, tel: 7891}

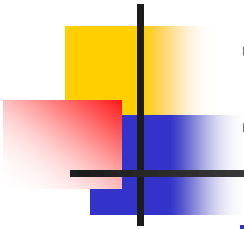
Introdução

- Representação Gráfica

nós representam objetos conectados por arestas que os descrevem

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

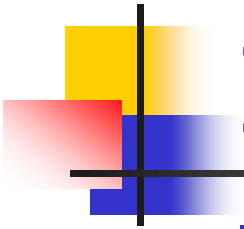




Introdução

- Situações Típicas

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



Introdução

- Exemplos
- Arquivos BibTex
- Têm estrutura, mas ela 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}
}
```

Introdução

- Exemplos
- Guia de restaurantes

```
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"
```

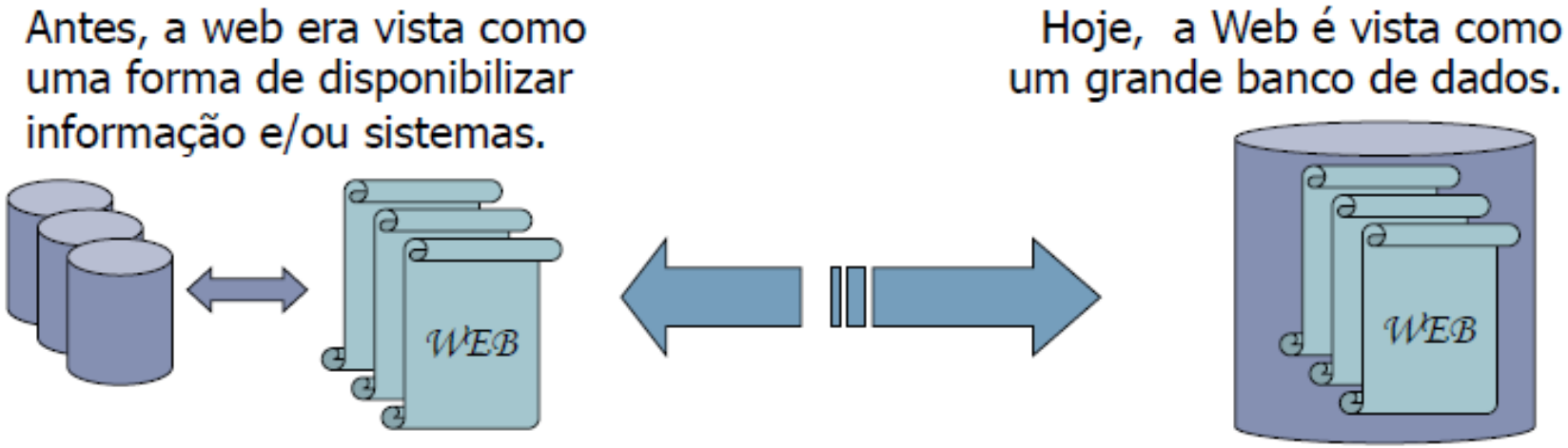
Introdução

- Exemplos
- Arquivos JSON

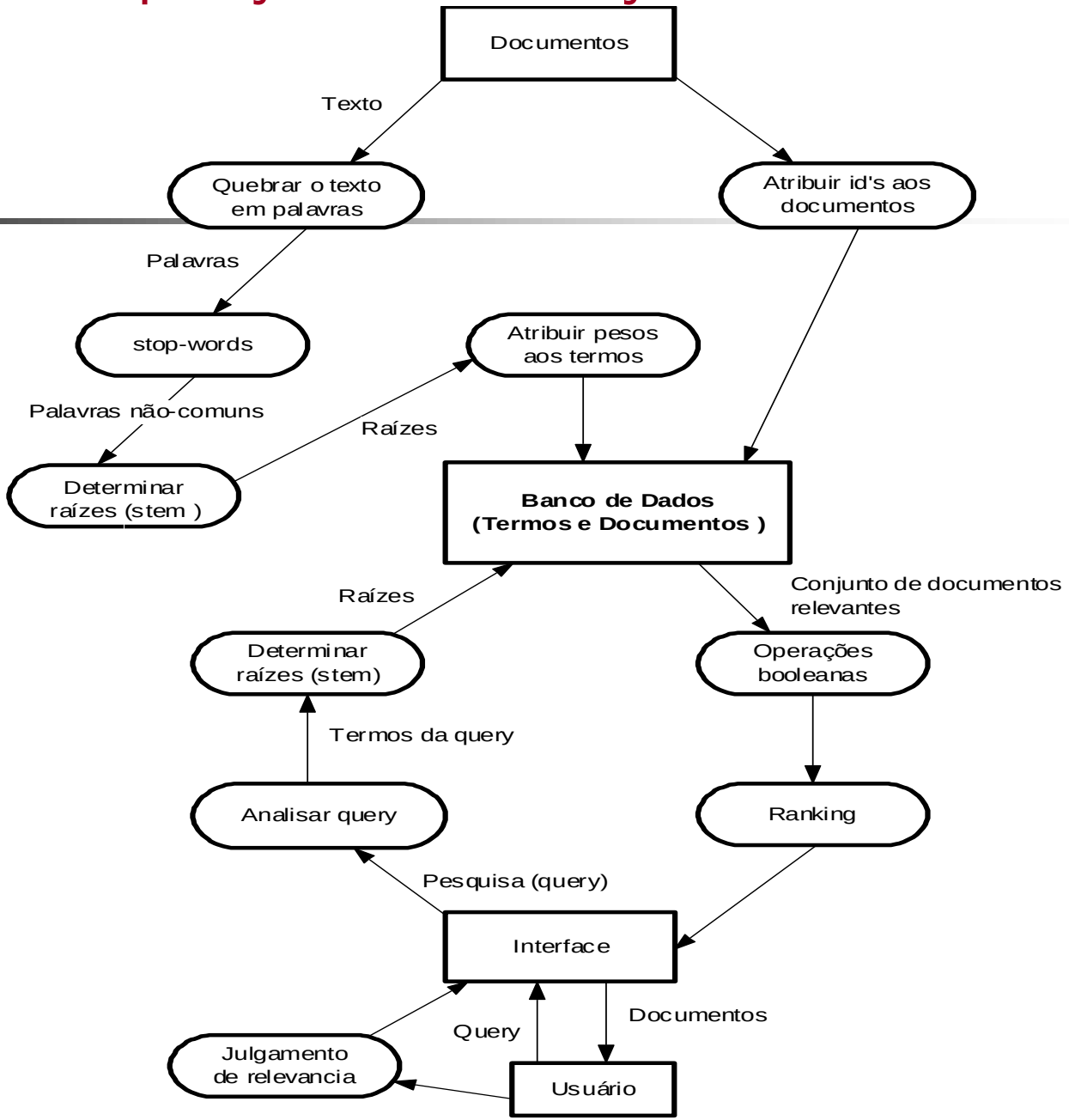
```
{ "Aluno" : [
   { "nome": "João", "nota": [ 8, 9, 7 ] },
   { "nome": "Maria", "nota": [ 8, 10, 7 ] },
   { "nome": "Pedro", "nota": [ 10, 10, 9 ] }
   ]
 }
```

Introdução

- Web e Dados Semi Estruturados

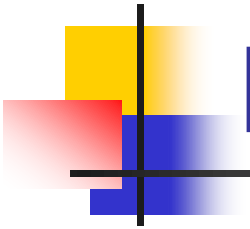


Sistema de Recuperação de Informação



Introdução

- **Parâmetros de qualidade da recuperação:**
  - **Revocação** (Retorno) R: total de itens significantes recuperados.
    - R = (número de itens relevantes recuperados) / (número de itens relevantes no sistema)
  - **Precisão** (Cobertura) P: total de itens significantes entre os recuperados.
    - P = (número de itens relevantes recuperados) / (número de itens recuperados)



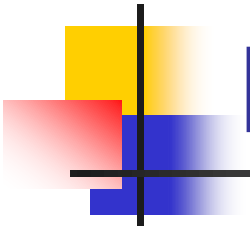
Busca na Web

- Problema:

  Como melhorar os resultados das nossas buscas na Web?

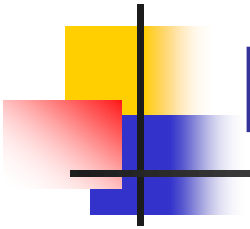
Tecnicamente:

Como maximizar Recall e Precision?



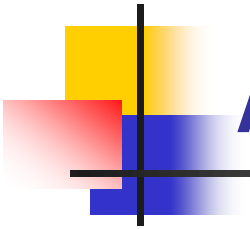
Busca na Web

- Desafios:
  - Escala
  - Diversidade
  - Mudança constante



Dados semi-estruturados ?

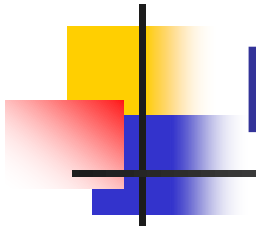
- Dados semi-estruturados: dados heterogêneos e irregulares, auto-descritivos.
- Banco de dados: originalmente, sistemas fechados, dedicados a gestão de dados regulares, cuja estrutura pouco evolui no tempo
- Novas aplicações requerem mais flexibilidade de representação e estão constantemente evoluindo o esquema
- Os modelos relacional e de objetos chegaram a seus limites (?)



A Web hoje...

- documentos HTML (em sua maioria)
- voltada para uso humano
- gerado automaticamente por aplicações
- fácil de alcançar qualquer Web page, de qualquer server, em qualquer plataforma

O Segredo do sucesso de HTML

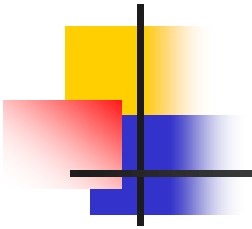


HTML é **simples**: todo mundo pode escrever HTML

HTML é **textual**: é legível, pode-se usar qualquer editor, ...

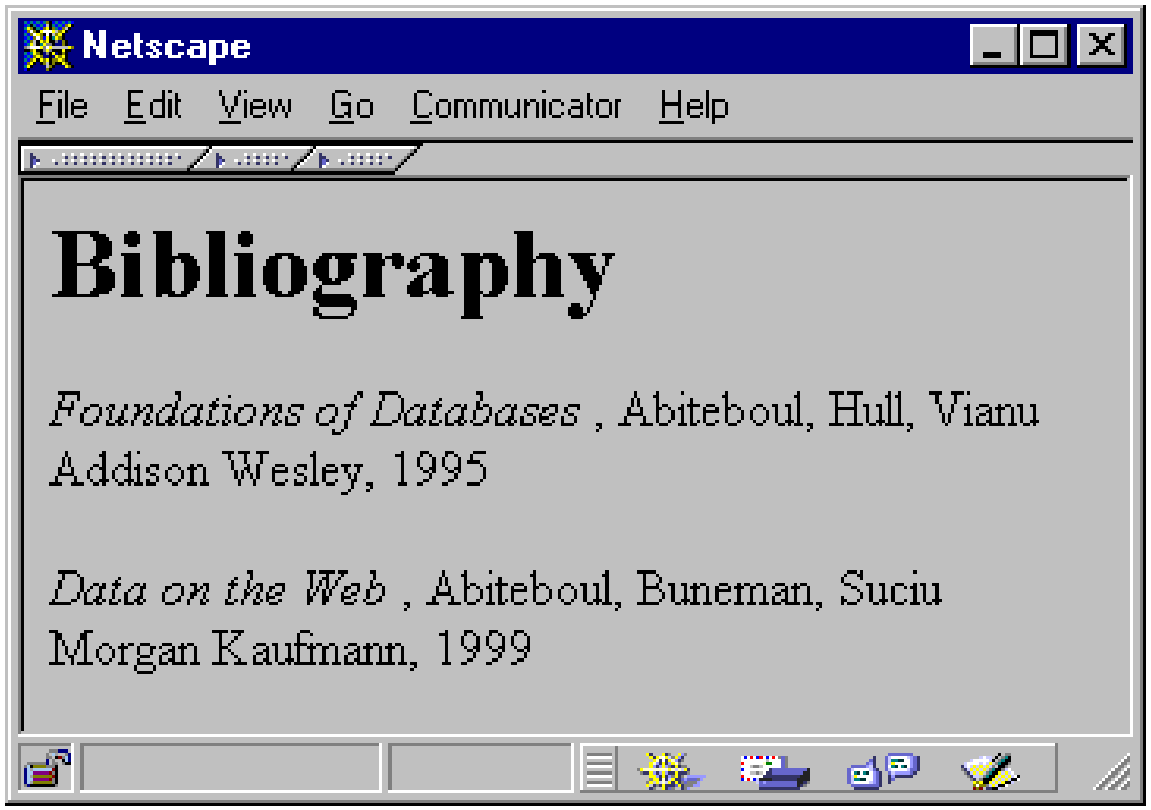
HTML é **transportável** em qualquer plataforma (o browser é a aplicação universal)

HTML conecta pedaços de informação através de **hypertext links**

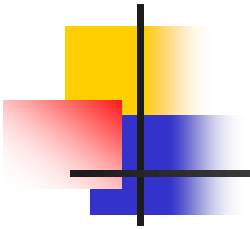


HTML

```
<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
```

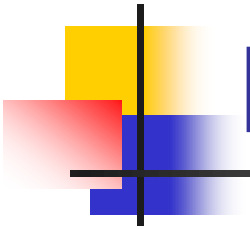


HTML descreve a apresentação



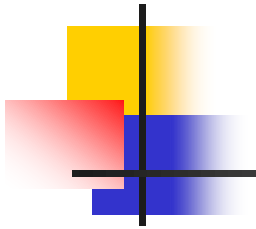
HTML ...

- Um conjunto pré-definido e limitado de tags, definidas por uma
- Estas tags possuem semânticas variadas:
  - `h1`,..,`h6`, `title`, `address`, ... dando as indicações estruturais
  - `center`,`hr`,`b`,`i`,`big`,`small`,... servem para descrever a apresentação.



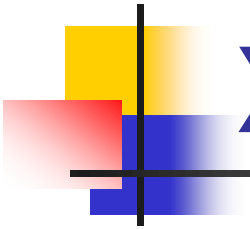
Limites da Web …

- aplicações não devem consumir HTML
- tecnologia de wrapper HTML é instável (modifica-se a página => modifica-se o wrapper)
- companhias se fundem, formam parcerias; necessitam de interoperabilidade de forma rápida



- As novas aplicações
  - Comércio Eletrônico
  - Protocolos "B2B"
  - Bibliotecas digitais
  - sistemas distribuídos
  - ...

precisamos de um "super HTML"



XML ?

- eXtensible Markup Language
- Uma linguagem de descrição de documentos, definida por um organismo internacional W3C
- Um conjunto de tecnologias derivadas
- O esperanto da Web

Web: Mudança de paradigma

...

- Novo padrão Web XML:
  - XML gerado por aplicações
  - XML consumido por aplicações
- troca de dados
  - entre plataformas: interoperabilidade na empresa
  - entre empresas

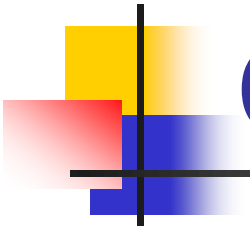
XML

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

XML descreve o conteúdo

XML: uma resposta única a necessidades variadas

- HTML é usada como a lingagem universal de apresentação de documentos na Web, mas não é uma linguagem adaptada para descrever *a estrutura* destes documentos
- Os sistemas de bases de dados atuais são muito rígidos para manipular dados cuja estrutura é *irregular* e evolui com o tempo.



Os segredos de XML

- Como HTML:
  - simples, legível, fácil de aprender
  - universal e transportável
  - suportado pela W3C (indústria absorve!)
- Mas, Além de HTML
  - flexível : podemos representar qualquer tipo de informação
  - estensível: pode-se representar informação de qualquer forma

Dois mundos se juntam na Web

Gestão de documentos

SGML

Documentação hipertexto

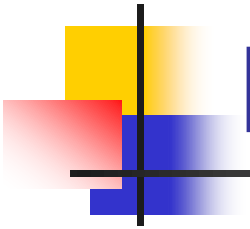
HTML

Gestão de dados

Bases de dados estruturados (relacionais e objeto)

Bases de dados semi-estruturadas

XML



Diferenças

- XML não é um substituto para HTML.
- XML e HTML foram projetados com objetivos diferentes:
  - XML foi projetado para transportar e armazenar dados, com foco no que os dados são;
  - HTML foi projetado para exibir dados, com foco em como os dados aparecem;
  - HTML é sobre a exibição de informações, enquanto o XML é sobre carregar informações.

Diferenças

```
<!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, 199

  </BODY>
</HTML>
```

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

Diferenças

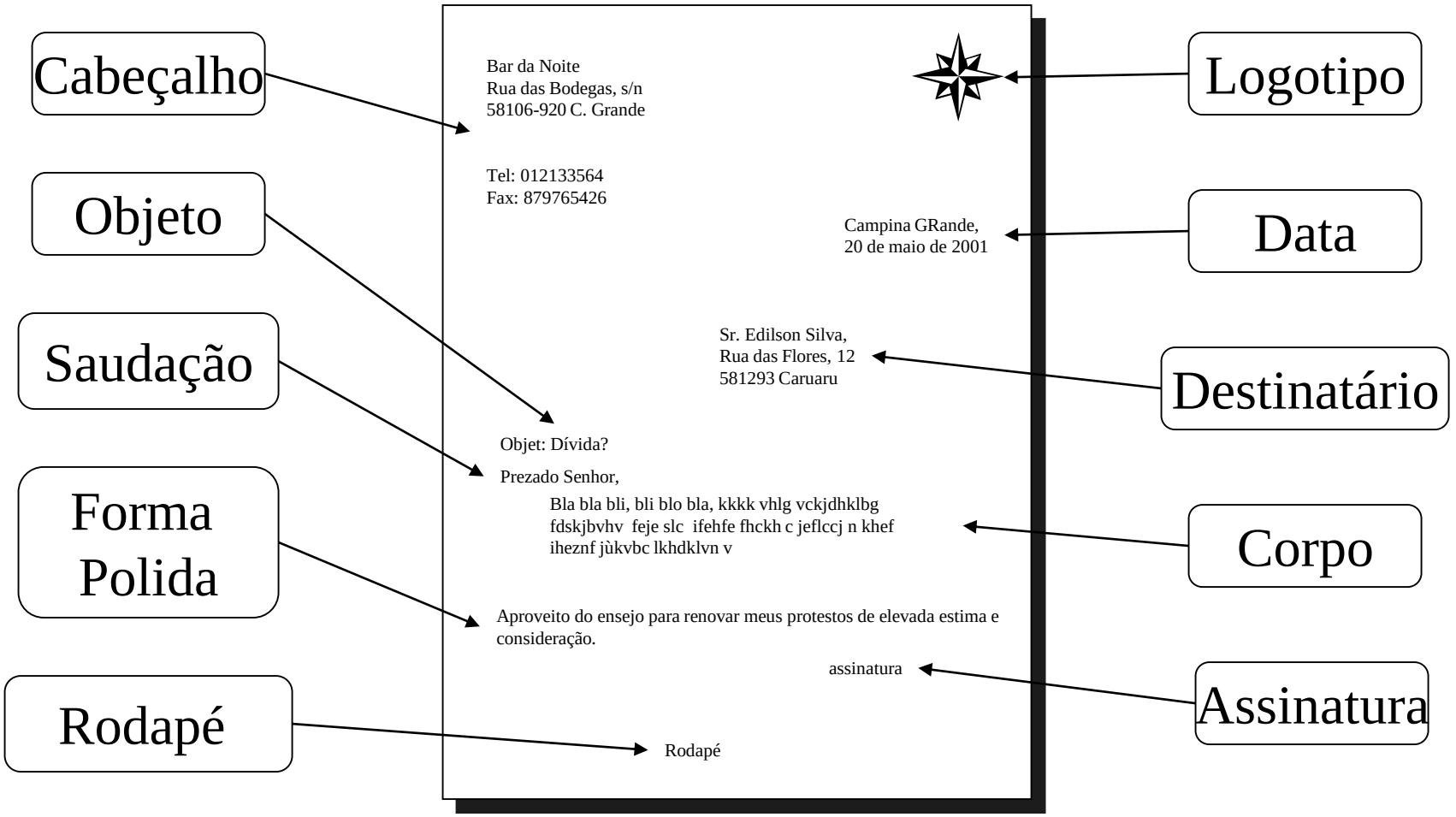
XML: Elementos (tags) definidos pelo usuário da linguagem e servindo para descrever o conteúdo e a estrutura.

Fonte XML

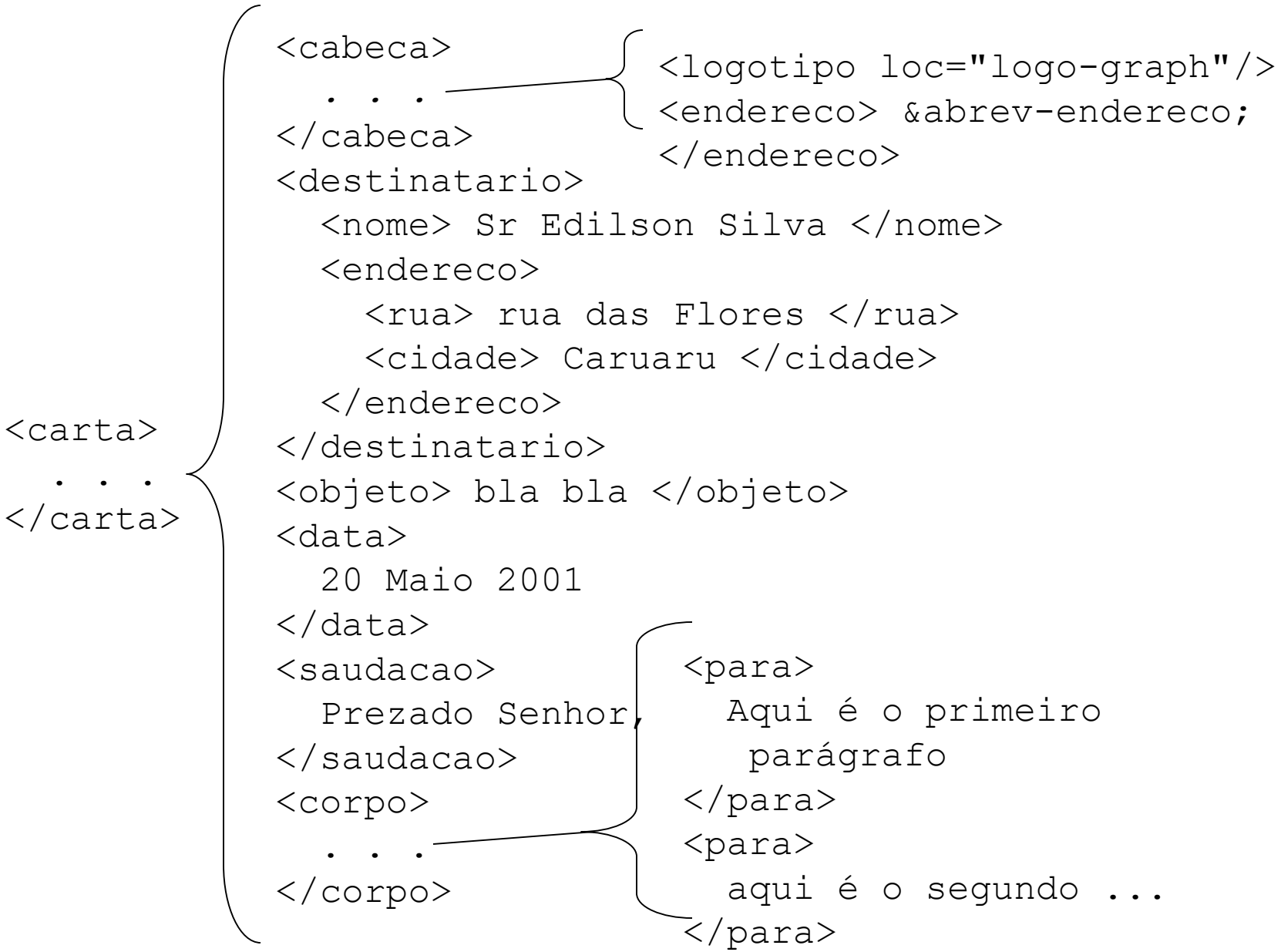
```
<bibliography>
    <book>    <title> Foundations... </title>
              <author> Abiteboul </author>
              <author> Hull </author>
              <author> Vianu </author>
              <publisher> Addison Wesley </
  publisher>
              <year> 1995 </year>
    </book>
    ...
</bibliography>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

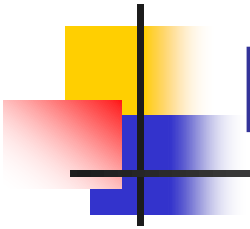
XML descreve o **conteúdo**!!!

Exemplo de um documento



Representação XML





Pontos importantes

- A representação desta carta em XML não tem nenhuma indicação sobre a apresentação.
- As numerosas propriedades gráficas ou tipográficas estão ausentes da fonte XML.
- Estas propriedades serão definidas por intermédio de uma *folha de estilo*.
- Uma folha de estilo é um *conjunto de regras* para especificar a *realização concreta* de um documento sobre uma *mídia* particular.

Consulta a dados semi-estruturados