



deti

universidade de aveiro
departamento de electrónica,
telecomunicações e informática

Engenharia de Dados e Conhecimento 2019/2020

Introdução aos Dados Semi-Estruturados



Dados Semi-Estruturados

- ▶ Em muitas aplicações, os dados não possuem um esquema rígido e pré-definido:
 - ▶ Ex: documentos estruturados, dados científicos, etc.
- ▶ A gestão deste tipo de dados requer outra forma de desenho dos componentes dos SGBDs:
 - ▶ modelo de dados, linguagem de pesquisa, otimização, sistema de armazenamento, etc.



Principais Caraterísticas

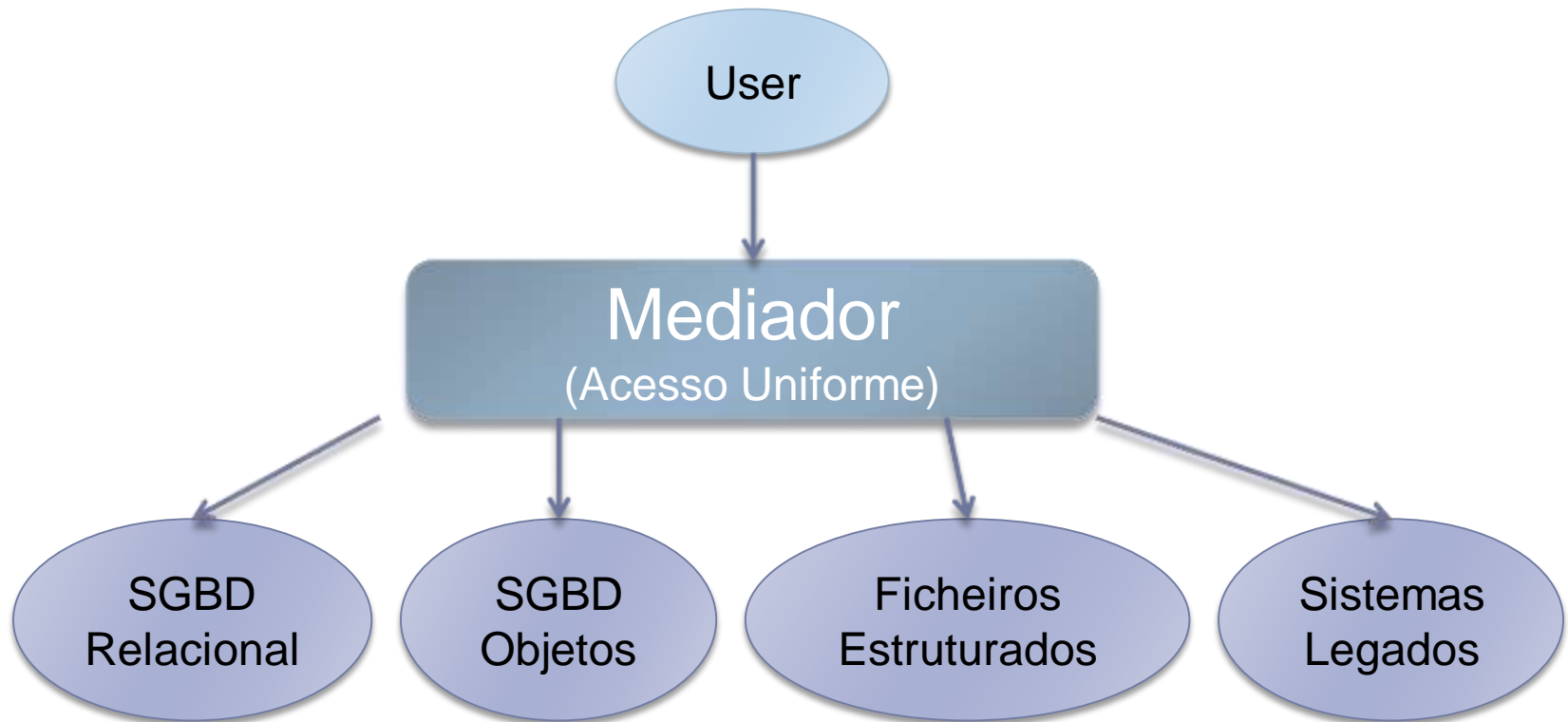
- ▶ O esquema não é o que costuma ser:
 - ▶ não é dado previamente
 - ▶ encontra-se muitas vezes implícito nos próprios dados
 - ▶ é mais descritivo do que prescritivo
 - ▶ é parcial
 - ▶ evolui rapidamente
 - ▶ pode ser muito grande
 - ▶ comparando com o tamanho dos dados
- ▶ Os tipos não são o que costumam ser:
 - ▶ objetos e atributos, não são fortemente tipificados
 - ▶ objetos, na mesma coleção, possuem representações diferentes



Problema 1 - Documento

```
<biblio>
  <livro ano="2004">
    <titulo>The Dark Tower</titulo>
    <autor>
      <apelido>King</apelido>
    </autor>
    <editora>Grant</editora>
  </livro>
  <livro ano="1991">
    <titulo>The Waste Lands</titulo>
    <autor>
      <nome>Stephen</nome>
      <apelido>King</apelido>
    </autor>
    <isbn>9788466300223</isbn>
  </livro>
</biblio>
```

Problema 2 – Integração de Dados



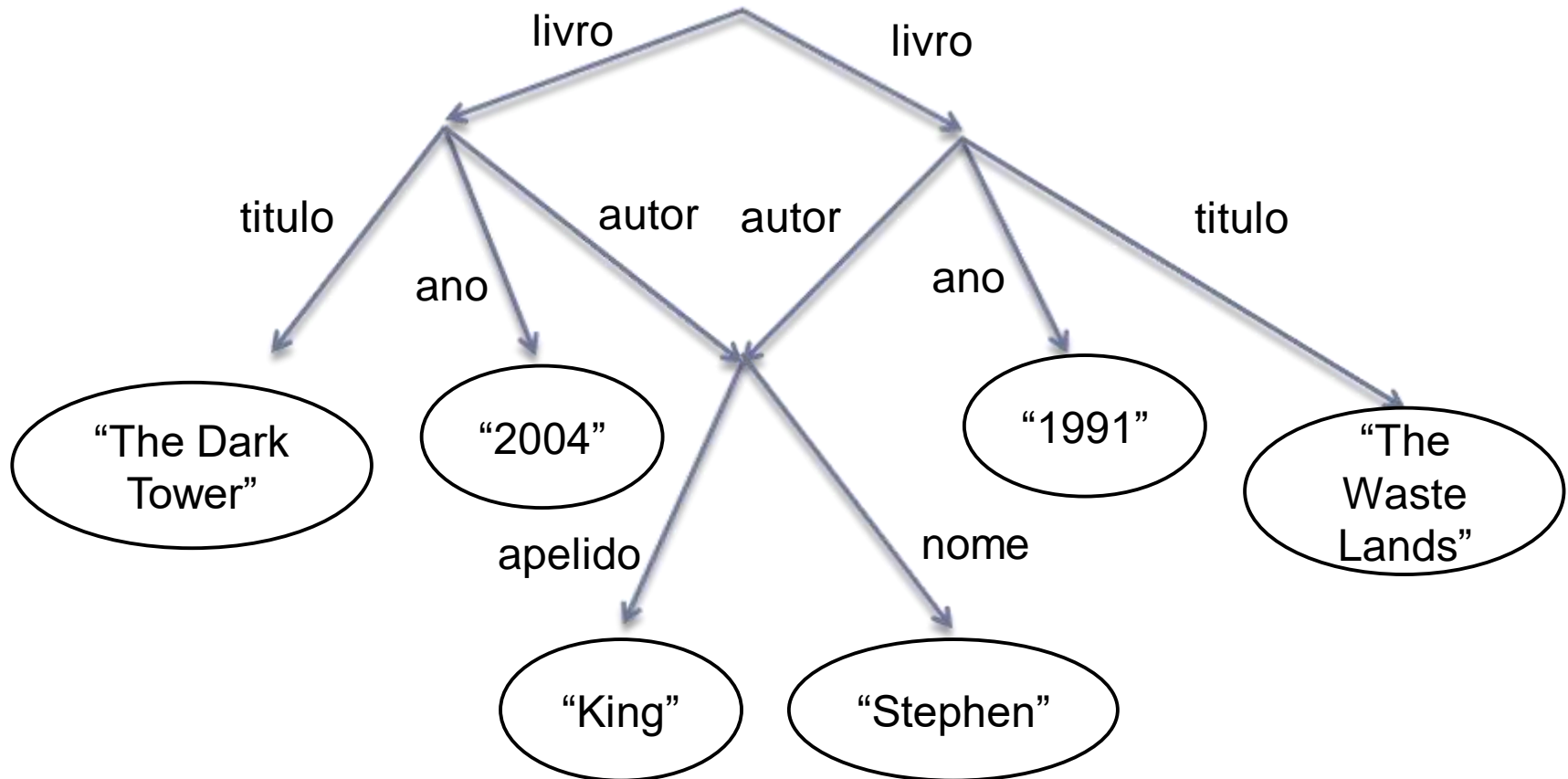
- ▶ Cada fonte representa os dados de forma diferente:
 - ▶ modelos e esquemas de dados diferentes

Modelos de Dados Semi-Estruturados



- ▶ Baseados em grafos
 - ▶ Os dados são armazenados nos nós
 - ▶ O esquema encontra-se nas ligações
- ▶ Usados para troca e integração de dados provindos de fontes heterogéneas
- ▶ Por vezes chamados de:
 - ▶ sem esquema
 - ▶ ou auto-descritivos

Grafos - Exemplo





Terminologia dos Grafos (i)

- ▶ Um grafo (orientado) $G = (N, E)$, consiste:
 - ▶ num conjunto de N nós (*nodes*)
 - ▶ e num conjunto de E ligações ordenadas (*edges*)
- ▶ Cada ligação em E consiste num par de nós (x, y) , onde x é a fonte e y o destino
- ▶ Um caminho de x_1 até x_n é uma sequência de ligações
 - ▶ $(x_1, x_2), (x_2, x_3), \dots, (x_{n-1}, x_n)$
- ▶ O tamanho de um caminho é o número de ligações existente no mesmo
- ▶ Um nó r consiste na raiz do grafo G se existe um caminho de r para qualquer outro nó em G
- ▶ Um ciclo é um caminho de um nó para ele próprio
- ▶ Um grafo sem ciclos é chamado de acíclico
- ▶ Um grafo é orientado se as suas ligações possuem direção



Terminologia dos Grafos (ii)

- ▶ Um grafo é enraizado se possui uma única raiz
- ▶ Uma árvore é um grafo enraizado G , no qual existe um único caminho da raiz para qualquer outro nó em G
- ▶ Um nó é uma folha se não é a fonte de qualquer ligação
- ▶ Os grafos podem conter nomes de nós e/ou nomes de ligações
- ▶ Num grafo com nomes de ligações $G = (N, E, FE)$, FE é uma função que atribui um nome a cada ligação
- ▶ Num grafo com nomes de nós $G = (N, E, FN)$, FN é uma função que atribui um nome a cada nó



Object Exchange Model (OEM)

- ▶ O modelo OEM original, usava apenas nomes de nós
- ▶ Atualmente, usa-se uma variante na qual as ligações também possuem nomes
- ▶ Um grafo de dados OEM é orientado, enraizado e possui nomes associados às suas ligações e folhas
- ▶ Os nomes das ligações consistem em *strings*
- ▶ Só as folhas possuem nomes que consistem em valores



Sintaxe OEM

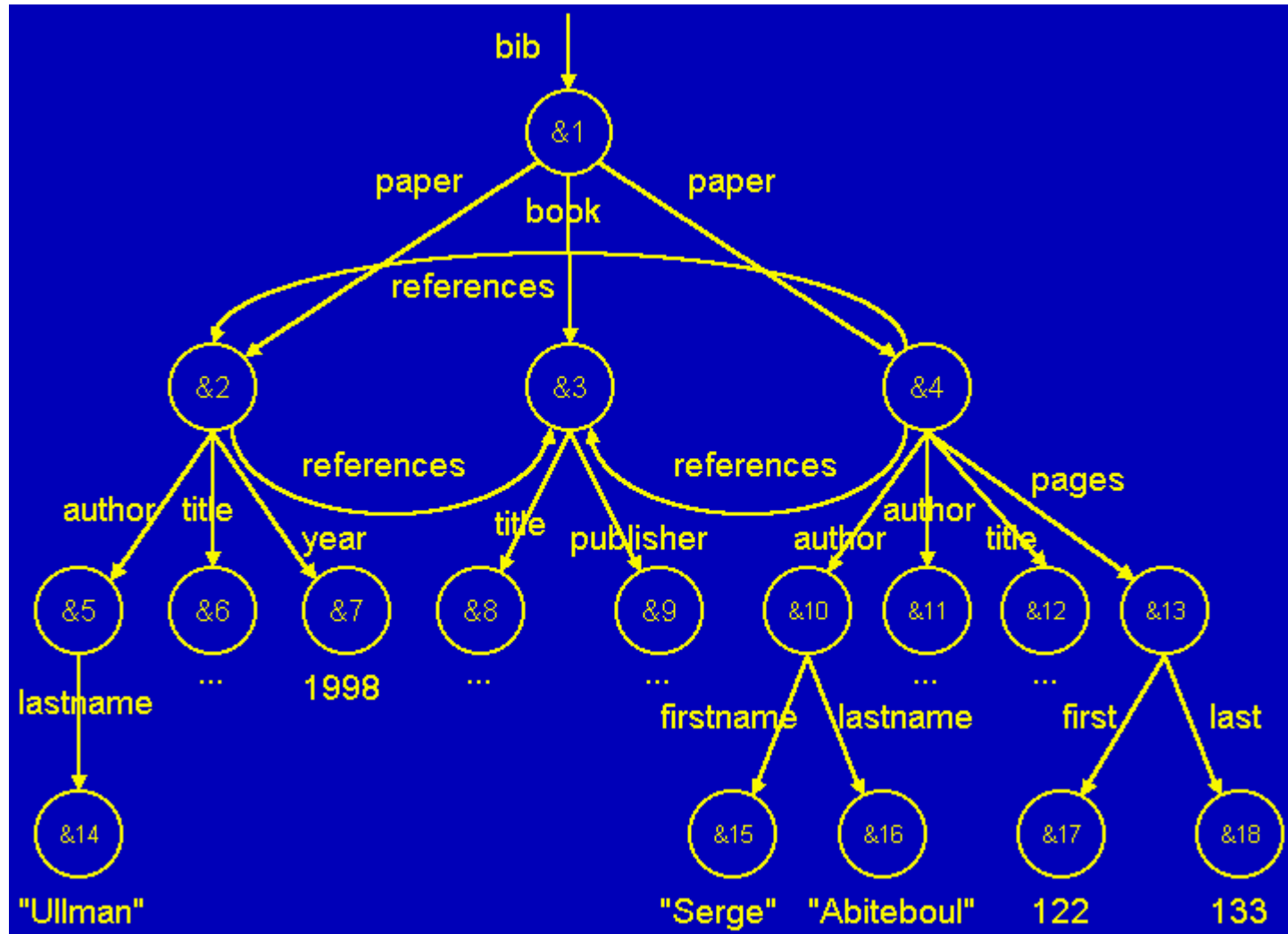
- ▶ Pares simples nome-valor (ou, chave-valor)
- ▶ Os nomes podem ser repetidos
 - ▶ Ex: múltiplos autores
- ▶ Para grafos que não são árvores
 - ▶ utilizar identificadores de objetos (ex: oids)

▶ Exemplo:

```
{ livro: { autor: { apelido: "King" },  
          titulo: "The Dark Tower",  
          ano: 2004  
        }  
}
```



Exemplo de Grafo OEM





Exemplo de Sintaxe OEM

```
bib: &1
{ paper: &2 { ... },
  book: &3 { ... },
  paper: &4
    { author: &10
      { firstname: &15 "Serge",
        lastname: &16 "Abiteboul" },
      author: &11 { ... }
      title: &12 { ... }
      pages: &13
        { first: &17 122,
          last: &18 133 },
      references: &2,
      references: &3
    }
  }
```

Avaliação do Modelo de Dados Semi-Estruturados



▶ Vantagens

- ▶ Fácil descoberta de dados novos e seu carregamento
- ▶ Fácil integração de dados heterogéneos
- ▶ Pesquisa fácil sem saber os tipos de dados

▶ Desvantagens

- ▶ Perca da informação de Tipo
- ▶ Otimização mais difícil



Esquemas de Grafos

- ▶ Dado um conjunto de dados semi-estruturados, é útil a extração de um esquema que o descreva
- ▶ Útil para:
 - ▶ procurar os dados pelos seus tipos
 - ▶ otimizar pesquisas, reduzindo o número de caminhos pesquisados
 - ▶ melhorar o armazenamento dos dados
- ▶ O esquema de um grafo especifica que ligações são permitidas num grafo de dados
 - ▶ todos os caminhos no grafo de dados ocorrem no esquema do grafo

Exemplo de um Esquema de Grafo

