

Importação de dados

Lendo dados de arquivo de texto pleno

Prof. Dr. **Walmes Zeviani** · UFPR

2021-05-06

Formatos de dados tabulares em texto pleno

Dados estruturados ou tabulares

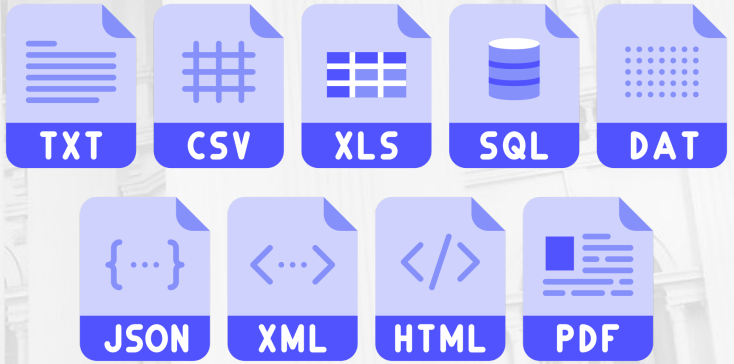
- São dados em formato tabular.
- São fornecidos como:
  - Arquivos de texto pleno.
  - Tabelas em planilhas eletrônicas.
  - Tabelas em BD relacionais.

Dados semiestruturados

- Possuem estrutura mas não é tabular.
- São fornecidos principalmente como:
  - Arquivos XML e derivados (HTML, KML, SVG, etc).
  - Arquivos JSON.

Dados não estruturados

- Não possuem estrura para armazenamento do conteúdo.
- Tipicamente são dados de texto pleno em
  - Documentos PDF.
  - Documentos HTML.



Fluxograma de uma estrutura de repetição tipo FOR. Fonte: os autores.

CSV · Valores separados por vírgula

- A sigla significa *comma separated values*.
- Os valores são separados em seus campos usando como delimitador a vírgula (`,`).
- Em português, usamos o delimitador `;` porque `,` é o separador decimal.

Dados de IDH e alfabetizados por município em 2000 extraídos do IPEA data.

```
Sigla;Código;Município;idh;renda;alfab;
AC;1200013;Acrelândia;;136,539;73,309;
AC;1200054;Assis Brasil;;115,16;70,954;
AC;1200104;Brasiléia;;132,383;75,493;
AC;1200138;Bujari;;118,815;61,23;
AC;1200179;Capixaba;;108,164;62,283;
AC;1200203;Cruzeiro do Sul;;140,199;71,548;
AC;1200252;Epitaciolândia;;134,364;75,381;
```

Dados disponíveis em http://leg.ufpr.br/~walmes/data/renda-idh-alfab.csv.

TSV · Valores separados por tabulação

- A sigla significa *tab separated values*.
- Os valores são separados em seus campos usando como delimitador tabulação.

Dados sobre resultados das corridas de F1.

```
circuito    data    rank    corredor    equipe    tempo    polip    pontos
Monza    03 September 1950    1    Giuseppe Farina    Alfa Romeo    2:51:17.400    3    8
Monza    03 September 1950    2    Dorino Serafini    Ferrari    -01:18.600    6    3
Monza    03 September 1950    2    Alberto Ascari    Ferrari    -01:18.600    6    3
Monza    03 September 1950    3    Luigi Fagioli    Alfa Romeo    -01:35.600    3    4
Monza    03 September 1950    4    Louis Rosier    Lago    -5 laps    13    3
Monza    03 September 1950    5    Philippe Etancelin    Lago    -5 laps    16    2
```

Dados disponíveis em http://leg.ufpr.br/~walmes/data/allgrandprix.txt.

FWF · Campos de comprimento fixo

- A sigla significa *fixed width format*.
- Os valores estão em campos com numero determinado de digitos.
- Então não há um caractere separador e precisa-se saber o comprimento dos campos.

Resultados da corrida de São Silvestre.

```
   1     2 PAUL TERGAT                (69)     M2529 0:44:47        FILA
   2     6 HENDRICK RAMAALA           (72)     M2529 0:45:05        REEBOK
   3     4 ELIJAH KIPTARBEI LAGAT     (98)     M0014 0:45:08        FILA KENYA
   4     7 SILVIO GUERRA              (68)     M3034 0:45:17        REEBOK ECUADOR
   5    12 JOHN GWAKO                 (78)     M2024 0:45:29        KENYA ADIDAS
   6     1 EMERSON ISER BEM           (73)     M2529 0:45:37        ASICS FUNILENSE SCS
   7   134 JOSE TELES DE SOUZA        (71)     M2529 0:45:56        HIDRAULY SESI STO ANDRE
   8    25 GERALDO DE ASSIS           (65)     M3034 0:46:04        ASE JUIZ DE FORA
   9    14 PATRICK NDAYISENGA         (71)     M2529 0:46:11        BURUNDI
```

Dados disponíveis em http://leg.ufpr.br/~walmes/data/saosilvestre_fwf.txt.


?
Dúvidas, comentários, sugestões?