University of Jijel
Faculty of Exact Sciences and Computer Science
Department of Computer Science
L3 – Computer Systems

Semi-Structured Data
Chapter 1
Introduction

Tarek Boutefara
t_boutefara@univ-jijel.dz
2025/2026

Content

- Structured data,
- Unstructured data (free format)
- Internet and Web
  - HTML documents
- Semi-structured data

Content

- **Structured data,**
- Unstructured data (free format)
- Internet and Web
  - HTML documents
- Semi-structured data

Structured data

- Databases
  - A database is a large collection of structured information stored on a permanent medium.

Structured data

- Databases
  - Overcoming the drawbacks of using files,
  - Centralized architecture,
  - Large "quantity" of data,
    - Static part of an information system,
  - Use via DBMS

Structured data

- Database Management Systems
  - A Database Management System (DBMS) is a high-level software application that allows you to manipulate the information stored in a database.

Structured data

- Database Management Systems
  - Unlike the file level (DBMS):
    - Open/Close
    - Read (binary array)/Write
  - A DBMS offers:
    - Data Manipulation Language (DML),
    - Data Definition Language (DDL),
    - Grant/Revoke

Structured data

- Database Management Systems
  - Record-based storage
    - A data structure for each table (file),
    - Binary files readable only by the DBMS,
  - And data exchange?
    - Dump?

Content

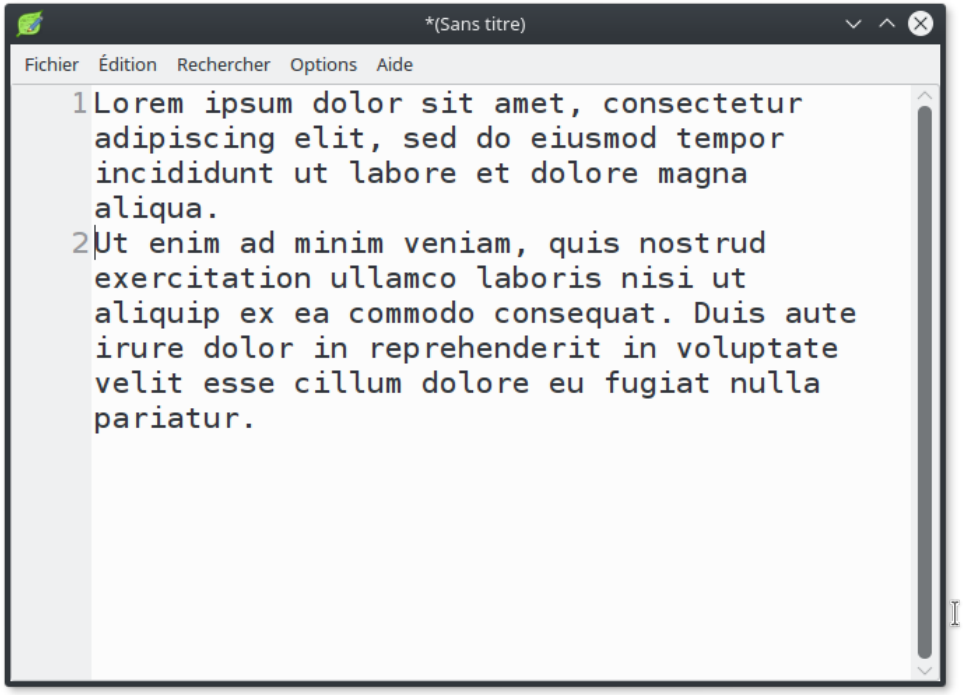
- Structured data,
- **Unstructured data (free format)**
- Internet and Web
  - HTML documents
- Semi-structured data

Unstructured data

- Free-format data:
  - Text files (most commonly used format):

Unstructured data

- Free-format data:
  - Text files (most commonly used format):



Unstructured data

- Open-format data:
  - Human-readable,
  - Difficult for machines to read and process.

Unstructured data

- C

*(Sans titre)

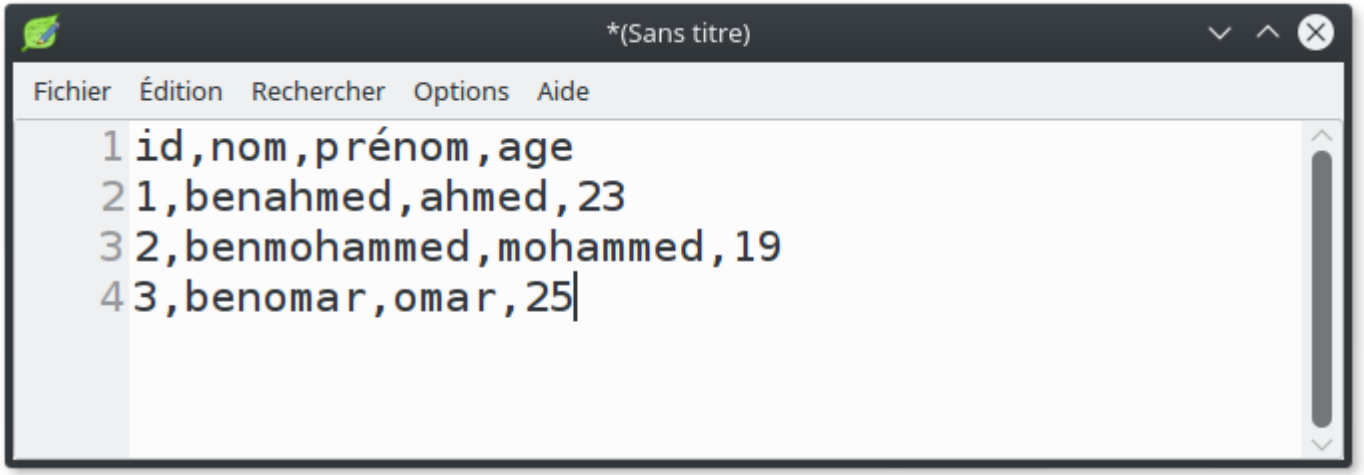
Fichier Édition Rechercher Options Aide

```
 1 Ingrédients
 2 5 personnes
 3
 4 Pain dur en morceaux    500 g
 5 Lait                    0.5 L
 6 Oeuf                    3
 7 Sucre vanillé           1 sachet
 8 Sucre en poudre         125 g
 9 Cannelle                1 pincée
10
11  Préparation
12
13 1 Faire tremper le pain coupé en morceaux
   dans le lait tiédi. Écraser à la fourchette
   quand le tout est bien ramolli.
14
15 2 Ajouter les œufs en omelette, le sucre,
```

ess.

Unstructured data

- Free-form data:
  - CSV files
    - Comma Separated Values
    - Structure?



Content

- Structured data,
- Unstructured data (free format)
- **Internet and Web**
  - HTML documents
- Semi-structured data

Internet and Web

- Internet
  - Global network,
  - Network of networks,
  - Communication infrastructure:
    - Packet transmission,
    - Routing,
    - Acknowledgment,
  - TCP/IP

Internet and Web

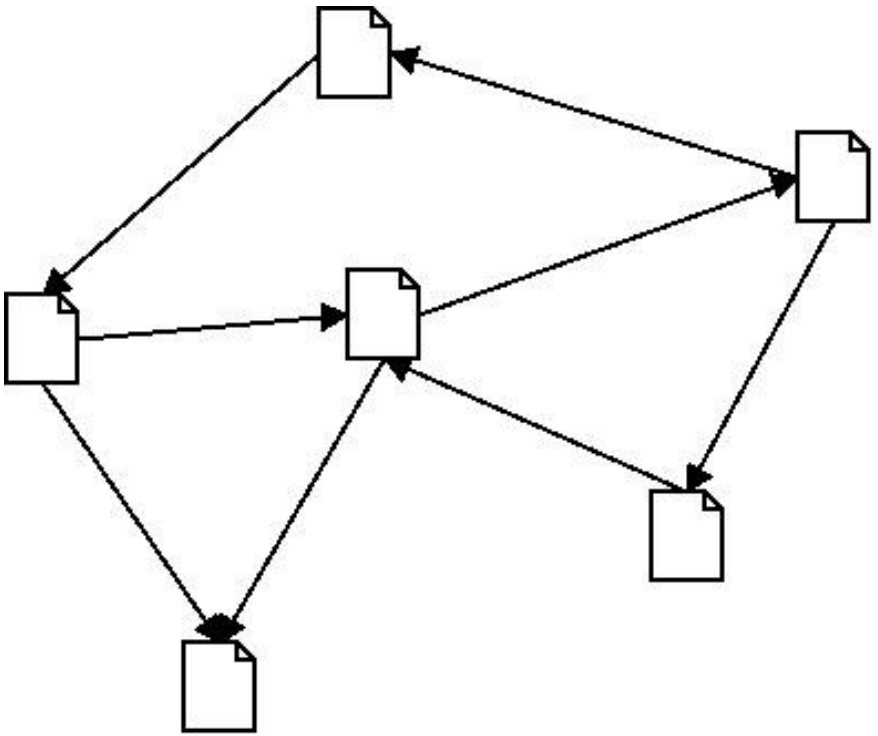
- Internet
  - Use cases:
    - Remote access
      - Telnet
      - Remote desktop
    - Data exchange
      - Email
      - FTP
      - Web

Internet and Web

- Web
  - An internet service,
  - Designed to allow the publication of content
  - With standardized access
    - All sites are accessible using the same tool
    - The web browser.

Internet and Web

- Web
  - Hypertext :
    - Non-linear :



Internet and Web

- Web
  - Intensive data exchange
    - Requests
    - Content
    - Authentication information

Internet and Web

- Web
  - Protocol: HTTP
    - Hyper Text Transfer Protocol
  - Language: HTML
    - Based on SGML (1986)
      - Standard Generalized Markup Language
    - Transfer via Document.

Content

- Structured data,
- Unstructured data (free format)
- Internet and Web
  - **HTML documents**
- Semi-structured data

Internet and Web

- Web
  - SGML:
    - Description language,
    - Separates logical structure from representation
      - Allows viewing of the document on different media
    - Introduction of the concept of style sheets.

Internet and Web

- Web
  - HTML:
    - Same principles (tags, stylesheets, etc.)
    - Creation of documents for viewing in the browser,
    - Hypertext support
      - "a database (text) format in which information related to that on a display can be accessed directly from the display"
    - Enhanced to support hypermedia

HTML Documents

- <!DOCTYPE html>
- <html>
  - <head>
    - <title>Example</title>
  - </head>
  - <body>
    - This is a paragraph.
  - </body>
- </html>

Content

- Structured data,
- Unstructured data (free format)
- Internet and Web
  - HTML documents
- **Semi-structured data**

Semi-structured data

- Semi-structured data is data that doesn't adhere to a fixed structure (like the tables in a relational database) but contains tags that separate different semantic elements.
- It falls somewhere between structured and unstructured data.

University of Jijel
Faculty of Exact Sciences and Computer Science
Department of Computer Science
L3 – Computer Systems

Semi-Structured Data
Chapter 1
Introduction

Tarek Boutefara
t_boutefara@univ-jijel.dz
2025/2026