Semi Structured Data

3rd year computer science (license)

2025/2026

Chapter 1
Context and Problem Statement

1.1. Review of Databases

Databases play a central role in information systems by enabling the storage, organization, and retrieval of data.

Traditionally, most applications have relied on **relational database management systems (RDBMS)**, where data is structured into tables composed of rows and columns and manipulated using SQL.

1.1. Review of Databases

Relational databases are well-suited for:

- highly structured and homogeneous data;
- applications with stable schemas;
- strong consistency and integrity constraints.

1.1. Review of Databases

However, with the evolution of applications, especially Web-based systems, several limitations have emerged:

- rigid schemas that are difficult to evolve;
- poor support for hierarchical or irregular data;
- limited ability to represent documents, multimedia content, and heterogeneous information.

These limitations motivated the exploration of alternative data models beyond the purely relational approach.

1.2. Multimedia and Digital Documents

A **digital document** is a structured or semi-structured collection of information intended for human or machine consumption.

Unlike relational records, documents may contain:

- textual content;
- images, audio, and video;
- metadata describing structure, semantics, or presentation.

1.2. Multimedia and Digital Documents

Multimedia data introduces additional challenges:

- large volumes of data;
- diverse formats and encodings;
- complex relationships between content and structure.

Traditional databases are not designed to naturally manage such data, particularly when structure is implicit or varies from one document to another.

1.3. Hypermedia, Internet, and the Web

The emergence of the Internet and the World Wide Web radically changed how data is produced, published, and consumed.

Web data is characterized by:

- distribution across heterogeneous systems;
- weakly structured or semi-structured formats;
- extensive use of links and references (hypermedia);
- frequent schema evolution.

1.3. Hypermedia, Internet, and the Web

Hypermedia systems combine documents with hyperlinks, enabling non-linear navigation between resources.

Standards such as HTML and later XML-based technologies were introduced to facilitate data exchange and interoperability across platforms.

These characteristics exposed the inadequacy of classical data models for Web-scale information management.

1.4. Problem Statement of Semi-Structured Data

Semi-structured data lies between structured relational data and completely unstructured data.

It is characterized by:

- the presence of structure, but not fixed or strictly enforced;
- self-describing data, where structure and data are often mixed;
- hierarchical and nested organization.

Examples include XML documents, JSON files, Web service messages, and configuration files.

```
<liste>
    <module code="m1">
            <lib>xml</lib>
            <coef>3</coef>
    </module>
    <module code="m2">
            <lib>java</lib>
            <coef>2</coef>
    </module>
    <etudiant id="e1">
            <nom>Addi</nom>
            <prenom>Kamel</prenom>
            <note ref="m1">15</note>
            <note ref="m2">12</note>
    </etudiant>
      <etudiant id="e2">
            <nom>Talbi</nom>
            <prenom>Larbi</prenom>
            <note ref="m1">17</note>
            <note ref="m2">14</note>
    </etudiant>
</liste>
```

1.4. Problem Statement of Semi-Structured Data

The main challenges addressed in this course are:

- how to model semi-structured data;
- how to validate and transform document-oriented data;
- how to query and store such data efficiently;
- how to integrate semi-structured data into Web and database applications.

1.4. Problem Statement of Semi-Structured Data

This chapter establishes the conceptual foundations necessary to understand why XML and related technologies emerged as core solutions for managing semi-structured data.