Semi Structured Data

3rd year computer science (license)

2025/2026

Chapter2
Documents and Multimedia Hyperdocuments

2.1. Documents

2.1.1. Introduction

A **document** is a coherent unit of information intended to be created, stored, transmitted, and interpreted as a whole.

In information systems, documents differ from traditional database records in that they:

- may contain heterogeneous data (text, images, metadata);
- often have an internal hierarchical structure;
- are designed primarily for readability and exchange rather than strict normalization.

2.1. Documents

Documents can be classified into

- structured,
- semi-structured,
- and unstructured documents.

Semi-structured documents explicitly describe part of their structure, making them suitable for automated processing while remaining flexible.

2.1.2. Modeling Specific Documents

Modeling a specific document consists of defining the internal organization of a single document type.

This includes:

- identifying structural elements (sections, paragraphs, headers, metadata);
- defining relationships between elements (containment, ordering);
- specifying optional and repeated components.

2.1.2. Modeling Specific Documents

Document modeling is typically hierarchical and tree-based.

Unlike relational modeling, it does not require a fixed schema shared by all instances, which allows natural representation of irregular or evolving data.

Examples of specific documents include:

- an academic article;
- a product catalog;
- a technical report.

2.1.3. Modeling Document Classes

A **document class** represents a family of documents sharing common structural characteristics.

Modeling a document class involves:

- abstracting common elements across multiple documents;
- defining constraints on structure and content;
- enabling validation and interoperability.

Document class modeling prepares the ground for schema-based approaches such as DTDs and XML Schemas, which will be studied later in the course.

2.2. Hyperdocuments

A **hyperdocument** extends the notion of a document by introducing explicit links between documents or between parts of the same document.

These links enable non-linear navigation and information discovery.

Key characteristics of hyperdocuments include:

- nodes representing documents or document fragments;
- hyperlinks representing semantic or navigational relationships;
- support for distributed resources.

2.2. Hyperdocuments

Hyperdocuments are fundamental to Web systems, where information is interconnected rather than isolated.

Technologies such as HTML and XML-based linking mechanisms enable the construction and management of hyperdocuments.

2.3. Multimedia Content

Multimedia content refers to information composed of multiple media types, such as text, images, audio, video, and animations.

Managing multimedia data raises several challenges:

- storage of large and heterogeneous data objects;
- synchronization between different media components;
- description and indexing of content using metadata.

2.3. Multimedia Content

In semi-structured data systems, multimedia elements are often accompanied by descriptive metadata encoded in structured or semi-structured formats.

This separation between content and description enables efficient processing, querying, and integration within information systems.