Structured-Document Processing Languages

First: Overview and Arrangements
What is this course about?
1.1 Basics of Structured-Documents
2.1 XML & XML Documents

Goals of the Course
- To get familiar with central models and languages for
  - manipulating
  - representing or styling
  - transforming
  - querying
  structured documents (or XML)
- "Generic XML processing technology"
  - very little about specific XML applications, or commercial systems

NOT an Exhaustive Survey
- Emphasis on manipulation of data in the form of documents, rather than describing/modelling data
- Bias in selecting course topics:
  - estimated usefulness/value
    - centrality (→ longer-lasting value)
    - maturity: Stable specifications?
    - Existing implementations?
  - Lecturer up-to-date?

Motivation?
- Academic interest in models of information processing
- Practical relevance of documents and XML
- Order → XML → Invoice
- Internet

Course Outline
1 Introduction
  Overview and Arrangements
  1.1 Structured Documents
2 XML Basics
  - XML documents, DTDs, Namespaces
3 Programmatic Manipulation of Structured Documents
  (XML APIs)
  3.1 SAX
  3.2 DOM
  3.3 JAXP

Course Outline (2)
4 Introduction to Styling Structured Documents
  4.1 Essentials of CSS
5 Transforming Structured Documents
  5.1 Addressing: XPath
  5.2 XSLT
6 Styling Structured Documents II: XSL
7 Querying Structured Documents
  - W3C XML Query Language XQuery

Methodological Goals
- Central professional skills
  - consulting technical specifications
  - experimenting with SW implementations
- Ability to think…?
  - to find out relationships, reason, explain, ...
  - to apply knowledge in new situations
- ("Pidgin English" for scientific communication)
We concentrate on textual or text-based documents
- character data major constituent of information content
- as opposed to, say multimedia documents

Next: Presentation vs Structure
### Markup

- **Procedural markup**
  - formatting commands (start boldface, produce an empty line, indent 5 mm, ...)
  - proprietary word processor formats, nroff, TeX, ...
- **Descriptive or generic markup**
  - indicating the logical structure of text using chosen names
  - LaTeX: \begin{abstract} \end{abstract}
  - HTML: <TITLE> ... </TITLE>
- **Markup language (merkkauskieli)**
  - a fixed set of markup notations (e.g. nroff, TeX, HTML, SVG, ...)

### Structured Documents?

Most liberally, *any* document is structured (punctuation, words, sentences, fields, ...) but especially descriptively marked-up documents ... (e.g. well-formed XML) especially if they adhere to a rigorous specification of structure (e.g. XML+DTD)

### Structure in Documents

- **Hierarchy or nesting** is ubiquitous
  - chapters of books, warnings in maintenance manuals, ...
- **Linear order** essential in prose documents
  - less important in documents representing data objects
- **Hypertext** and **cross-references**
- We'll be mainly dealing with manipulation of hierarchical, or tree-like document structures

Next: How are these modelled?