Structured-Document Processing Languages

(3 cu), Spring 2003
Pekka Kilpeläinen
University of Kuopio
Department of CS & Applied Math
Pekka.Kilpelainen@cs.ukk.fi

1 Introduction

First: Overview and Arrangements
What this course is about?

1.1 Structured Documents
Review of basic notions

Goals of the Course

- To familiarize with the most important models and languages for
  - manipulating
  - representing
  - transforming and
  - querying
structured documents (or XML)
- "Core XML processing technology"
  - very little about specific XML applications or commercial systems

NOT an Exhaustive Survey

- Emphasis on active formalisms (for describing processes on documents) instead of describing documents/data
- Bias in selecting course topics:
  - estimated usefulness/value
    - centrality (implying longer-lasting value)
    - maturity: Stable specifications? Existing implementations?
  - Lecturer up-to-date?

Motivation?

- Academic interest on models of information processing
- Practical relevance: "eBusiness" is HOT!

Preliminary Outline

1 Introduction
  Overview and Arrangements
  1.1 Structured Documents
2 Document Instances and Grammars
  2.1 Trees and their Grammars
  2.2 Review of XML basics: DTDs, Namespaces, Schemas
3 Programmatic Manipulation of Structured Documents (XML APIs)
  3.1 SAX
  3.2 DOM, 3.3 JAXP
Preliminary Outline (2)

4 Styling Structured Documents I
   4.1 Essentials of Cascading Style Sheets
5 Transforming Structured Documents
   5.1 Addressing: XPath
   5.2 XSLT
6 Styling Structured Documents II: XSL
7 XML wrapping (or translating data to XML)
8 Querying Structured Documents
   - W3C XML Query Language XQuery

Methodological Goals

- Some central professional skills
  - consulting of technical specifications
  - experimenting with SW implementations
- Ability to think… ?
  - to find out relationships
  - to apply knowledge in new situations
- ("Pidgin English" for scientific communication)

Administration

- An elective graduate-level (laudatur) special course
  - suitable for all specialisation lines (esp. CS/SWE)
- 3 cu (=120 hours of work)
- Lectures Jan 15 -, Microteknia MT2
  - Lecturer: Pekka Kilpelainen@uku.fi
  - Assistant: Tommi.Penttinen@uku.fi

Administration: Exercises

- Exercises Jan 27 - March 10, MT2
  - essential for familiarizing with the technology
  - mainly normal homework assignments, solutions discussed in class
- + a few (1-2) "mini-projects"
  - reading and summarising tasks?
  - hands-on experimentation?
  - to be handed in to lecturer
  - credited like other exercises (scaled based on quality by a factor in [0, 1.5])

Administration: Grading

- Course examination on Thu, March 13, in Lecture hall L2
  - minimum of 50% of exam points to pass the course
- Grade =
  - (3×Exam+MaxExam + 12×HomeWork/MaxHomeWork - 8)/3
- Opportunity to retake the exam
  - April 3 (> 50% to pass; grade with/without homework credits, whichever is better)

Material

- No single textbook
- Reports, articles
- Course home page
  - http://www.cs.uku.fi/~kilpelai/RDV03/
    - lecture notes, exercises, reference material, announcements, ...
- Recommended (but not required) text:
Background Check

- Basic knowledge of structured documents and document standards
  - Course "Document standards"?
  - HTML?
- Programming languages and concepts
  - OO programming, Java?
  - Unix/Linux, Windows?
- Formal language theory
  - Theory of Computation / "Onyalmaarin ja lastiennan teoria"?
  - Regular expressions, automata?
  - Context-free grammars, parse trees?

Course Expectations?

1.1. Structured Documents

- Document:
  - A structured representation of information on some medium (= message)
  - Normally for a human reader
  - Also application-to-application messages
  - EDI (electronic data interchange)
  - "Prose-oriented XML" vs "data-oriented XML"
  - Possibly non-permanent, dynamically generated
  - Processable or conceivable as a unit
  - (a web page vs a web site)

Text-Based Documents

- 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

Presentation vs Structure

- Presentation informs the human reader about the meaning of text and the role of its parts
- Markup: indicating the presentation or the meaning of different parts of text
  - Originally hand-written annotations for the typesetter
  - Nowadays primarily codes embedded in digital documents

Markup

- Procedural markup
  - Formatting commands (start boldface, produce an empty line, indent 5 mm, ...)
  - Proprietary word processor formats, roff, TeX, ...
- Descriptive or generic markup
  - Indicating the logical structure of text using chosen names
  - LaTeX \begin{abstract} \end{abstract}
  - HTML: \section{Introduction} ...
- Markup language
  - A fixed set of markup notations (e.g. roff, TeX, HTML, SVG, ...)
Structured documents?

Most liberally, any document is structured (punctuation, words, sentences, fields, ...) but especially descriptively marked-up documents ... especially if they adhere to a rigorous specification of structure.

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 these are modelled?