1 Introduction

First: Overview and Arrangements

What is this course about?

1.1 Structured Documents

Review of basic concepts

NOT an Exhaustive Survey

- Emphasis on processing data in the form of documents, rather than describing it
- Bias in selecting course topics:
  - estimated usefulness/value
    - centrality (implying longer-lasting value)
    - maturity: Stable specifications?
    - Existing implementations?
  - Lecturer up-to-date?

Goals of the Course

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

Methodological Goals

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

Motivation?

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

Course 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
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

Bias in selecting course topics:

- estimated usefulness/value
  - centrality (implying longer-lasting value)
  - maturity: Stable specifications?
  - Existing implementations?
  - Lecturer up-to-date?

Practical relevance:

- XML processing
  - Academic interest in models of information processing
- "eBusiness" is HOT!
  - "eBusiness" for commercial systems

Order

Internet

XML invoice

Structured-Document Processing Languages

(3 cu), Spring 2005
Pekka Kilpeläinen
University of Kuopio
Department of Computer Science
Pekka.Kilpelainen@cs.uku.fi
Administration

- An elective graduate-level (laudatur) special course
  - suitable for all specialisation lines (esp. CS/SWE)
- 3 cu (=120 hours of work)
- Lectures: April 9 – May 25, MT2 (E26–27)
  - Lecturer: Pekka.Kipelainen@uku.fi
- Assistant: Tommi.Penttinen@uku.fi

SDPL 2005 Notes 1: Introduction 9

Administration: Exercises

- **Exercises** April 21 – May 26, MT2/S21
  - essential for learning the technology
  - mainly normal homework assignments, some hands-on practice; Solutions discussed in class
- **Course Project**
  - extending a document processing application (XML/Java/DOM/JAXP/XSLT)
  - individually or in small groups
  - solutions to be handed in to lecturer
- 50% of exam points required to pass the course
  - credited like other exercises, points ~3 ex. sessions (graded by a quality-based factor in [0, 1.5])

SDPL 2005 Notes 1: Introduction 10

Administration: Grading

- Final **exam** on Tuesday, May 31, in SL
  - ≥ 50% of exam points required to pass the course
- **Grade** =
  \[
  \text{floor}(12 \times \text{Exam/MaxExam} + 4 \times \text{HomeWork/MaxHomeWork} - 3)
  \]
- Opportunity to retake the exam
  - June 15 (again ≥ 50% to pass; grade with/without homework credits, whichever is better)

SDPL 2005 Notes 1: Introduction 11

Material

- No single textbook
- Reports, specifications, articles
- Course home page
  - http://www.cs.uku.fi/~kilpelai/RDK05/
  - slides, exercises, reference material, announcements

SDPL 2005 Notes 1: Introduction 12

Background Check

- Basic knowledge of structured documents and document standards
  - Course "Introduction to Document standards?"
  - HTML?
- Programming languages and concepts
  - Java? OO programming?
  - Unix/Linux vs. Windows?
- Formal language theory
  - Theory of Computation / "Ohjelmoinnin ja laskennan teoria"?
  - regular expressions, automata?
  - context-free grammars, parse trees?

SDPL 2005 Notes 1: Introduction 13

Course Expectations?

- No expectation of previous knowledge of XML
- rented course

SDPL 2005 Notes 1: Introduction 14

1.1. Structured Documents

- **Document:**
  - a structured representation of information on some medium (= message)
  - normally for a human reader
    - memos, manuals, articles, books, ...
  - 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)

SDPL 2005 Notes 1: Introduction 15

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 (merkkaus) indicates 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, 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?