XML Processing Techniques

Petrozavodsk State University
28.2. - 4.3.2006 (1 + 2 ECTS)
Prof. Pekka Kilpeläinen
Univ of Kuopio, Dept of Computer Science
Pekka.Kilpelainen@cs.uku.fi

Introduction

First: Overview and Arrangements
What is this course about?

- "Generic XML processing technology"
  - techniques applicable to arbitrary XML data
    - » APIs for programmatic manipulation
    - » XSLT for document transformations

Goals of the Course

- Learn about models and languages for
  - manipulating and
  - transforming

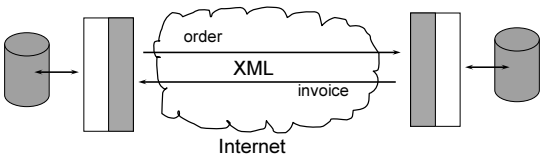
  XML data/documents, to be able
  - to use references
  - to learn more, and
  - to apply the technology

NOT an Exhaustive Survey

- Short version of a more comprehensive course
- Emphasis on **processing** data in the form of documents, rather than describing it
- Bias in selecting course topics:
  - estimated usefulness/value
    - » centrality (implying longer lifespan)
    - » maturity: Stability of specifications?
      Existence of implementations?

Motivation?

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



Course Outline

Intro and Arrangements; Structured documents & markup
1 Document Instances and Grammars
1.1 XML and XML docs; 1.2 Document grammars
1.3. XML DTDs; 1.4 XML Namespaces
2 Programmatic Manipulation of XML (XML APIs)
2.1 SAX; 2.2 DOM; 2.3 JAXP
3 Transforming XML
Overview and main aspects of XPath and XSLT
3.1 Additional features; 3.2 Computing with XSLT

Arrangements

- Lectures and exercises 1 ECTS cp
  (additional 2 cp by an optional course project)

Lectures		Exercises
Tue 28.2.	13.30-15.05, 15.15-16.50	
Wed 1.3.	13.30-15.05	15.15-16.50
Thu 2.3.	9.45-11.20	12.25-14.00
Fri 3.3.	13.30-15.05	15.15-16.50
Sat 4.3.	9.45-11.20	11.30-13.05

Course Project

- **Course Project (optional)**
  - extending a document processing application (XML/Java/DOM/JAXP/XSLT)
  - individually or in small groups
  - solutions handed-in to lecturer by March 17
  - instructions available at
    www.cs.uku.fi/~kilpelai/XPT06/project.html

Source Material

- No single textbook
  - Possibly useful background text: *Deitel, Deitel, Nieto, Lin & Sadhu: XML - How to Program. Prentice Hall, 2001.*
- Reports, specifications, articles
- Course home page at Univ. of Kuopio
  - `www.cs.uku.fi/~kilpelai/RDK05/`
  - slides, exercises, reference material

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
    - » e.g., btw client and server in **Web Services**
  - "prose-oriented XML" vs "data-oriented XML"
  - can be treated as a single unit
    - » (a web page vs a web site)

Presentation vs Structure

- Presentation informs the *human reader* about the meaning of text and the role of its parts
- **Markup** 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; **`<Tags>`**

Markup and Markup Language

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

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