PB138 — Introduction to XML

Outline

- Introduction to XML
- Motivation
- Principles
- Resources @FI (courses) and elsewhere

What is XML?

- XML is a standard by the W3C (http://www.w3.org) consortium prescribing how to create markup languages.
- It is therefore a *metalanguage*.
- It is ideologically based on older standards (SGML *Structure Generalized Markup Language*) — XML can be seen as almost a subset of SGML.
- There are several other standards closely related to XML, such as *XML Namespaces*, XInclude, XML Base, XML Infoset.
- These standards together with others (*XSLT*, *XSL-FO*, *XHTML*, *CSS* ...) form a "family" of XML standards.

Ten principles for the XML standards

From the preamble for XML 1.0 (Third Edition)

1. XML shall be straightforwardly usable over the Internet. XML bude přímočaře použitelné na Internetu.
2. XML shall support a wide variety of applications. XML bude podporovat širokou škálu aplikací.
3. XML shall be compatible with SGML. XML bude kompatibilní se SGML.

Ten principles for the XML standards (contd)

4. It shall be easy to write programs which process XML documents. Tvorba programů zpracovávajících XML bude jednoduchá.
5. The number of optional features in XML is to be kept to the absolute minimum, ideally zero. Počet volitelných prvků XML standardu bude málo, optimálně nula.
6. XML documents should be human-legible and reasonably clear. XML dokumenty by měly být "lidsky" čitelné a rozumně jednoduché.

Ten principles for the XML standards (contd)

7. The XML design should be prepared quickly. Návrh XML standardu by měl být rychle hotov.
8. The design of XML shall be formal and concise. Návrh XML musí byt formální a správný.
9. XML documents shall be easy to create. XML dokumenty bude možné snadno vytvořit.
10. Terseness in XML literal is of minimal importance. Úspornost XML značkování není podstatná

Characteristics of XML languages

- XML is not a specific markup language, it's a specification determining how the markup languages should look like,
- so it is a "meta-language",
- conceptually a simplification of the SGML standard to facilitate the creation of parsers (analyzers) and applications.
- As each element in an XML document must be closed, the documents need not have a DTD for structure recognition.

Characteristics of XML languages

- XML builds on a successful implementation of SGML - HTML. It has similar characteristics in terms of the focus on the Internet.
- Serious discussions are held around binary XML, which should be equivalent representations of the same model as the "text" XML.

Current specifications of XML

- original specification (W3C Recommendation) to the W3C XML 1.0: http: //www.w3.org/XML
- 5th Edition at Extensible Markup Language (XML) 1.0 (Fifth Edition) (http: //www. w3. org/TR/REC-xml)
- XML is 20 - the first spec is now 21 years old (Feb 10, 1998)
- XML 1.1 (Second Edition) (http://www.w3.org/TR/xmlll) - changes induced by the introduction of UNICODE 3, easier normalization, the specification of handling procedure for "end of line" characters . XML 1.1 is not bound to specific version of UNICODE, but always on the latest version.

W3C Activities

- XML Coordination Group intermediate-working group, kind of "interface" between different groups of activities and also externally

- XML Core Working Group development of major specifications (XML) and closely related ones (Namespaces in XML, XML Information Set, XInclude)

W3C Activities

- Efficient XML Interchange Working Group development of standards for effective exchange of XML data with emphasis on portability and platform independence of the individual products (including eg Binary XML Characterization)
- XML Processing Model Working Group working on the definition of a scripting language for XML, the specification operations over XML data
- XML Linking Working Group the now defunct group worked on the development of XML Linking Language XLink) and XML Pointer Language (XPointer).

W3C Activities

- XML Query Working Group is designing the XML Query Language (XQuery and XPath - together with XSL Working Group)
- XML Schema Working Group Prepares specifications of W3C XML Schema to describe the structure, content, or semantics of XML documents.

What next?

- Neither XML is an "ultimate solution" to all problems of machine data exchange. Development goes on.
- For interactive (rich) web applications (RIA) with intensive server-to-client communications, because of easier interpretability and smaller data, the formats such as JSON (JavaScript Object Notation) are used.
- YAML is used for handwriting structured data.
- These standards will be mentioned during lectures as well. The focus of the course is in XML, derived formats instruments for processing and applications.

Tutorials and papers

- Zvon XML Tutorial: http://www.zvon.org/xxl/XMLTutorial/General/
- Tutorial ke XML na W3 Schools: http://www.w3schools.com/xml/default.asp
- Microsoft XML Tutorial: http://msdn.microsoft.com/xml/tutorial/
- 101 XML Tutorials: http://www.xmll01.com/xml/default.asp
- XML Tutorials at Beginners.co.uk: http://tutorials.beginners.co.uk
- Tutorials at Developerlife.com: http://developerlife.com

Portals on XML

- World Wide Web Consortium (W3C)
- XML Startkabel
- Zvon — excellent collection of tutorials, on-line references in many languages, hosted in CZ
- XML Cover Pages — daily updated collection of links to articles, standards, software, etc. in XML. Best in this category.
- OReilly XML.COM — articles, tutorials at a high level
- IBM DeveloperWorks, section XML — papers, tutorials, software atd. at a high level

More links to XML

- Activities of W3C: http://www.w3.org/XML/Activity - specification of standards, conferences, links to SW, reference tools, links
- What is XML? na XML.COM: http: //www.xml.com/pub/a/98/10/guide0.html - one of the intro articles to XML
- XML: XML Quick Syntax Reference Card (http://www.mulberrytech.com) - great, simple reference card
- Commented version of XML specification at XML.COM (Annotated XML): http://www.xml.com/pub/a/axml/axmlintro.html

Books

- XML in Nutshell by E.R.Harold

Resources on XML at FI / Courses — Fall term

- PA165 Enterprise Java - T. Pitner, P. Adámek, M. Kuba, B. Rossi, F. Nguyen, M. Cupák, M. Brišk ár
- PB029 Electronic document preparation - P. Sojka
- PV110 Software electronic publications I - P. Sojka
- PV173 Seminary of NLP Lab

Courses — Spring term

- IB047 Intro to corpus linguistics and computer lexikography - K. Pala, P. Rychlý
- PA105 Technologies of Information Systems II - J. Král
- PA154 Corpus Tools - P. Rychlý

- PA156 Dialogue System - I. Kopeček
- PV174 Lab of Electronic and Multimedia Apps - P. Sojka
- PV030 Textual IS - P. Sojka
- PV113 SW electronic publications II - P. Sojka