PB138 — XML Information Set, Canonical Form, Security

XML Information Set

- *XML Information Set (Second Edition) W3C Recommendation*
- First published on 24 October 2001, revised 4 February 2004, John Cowan, Richard Tobin, http://www.w3.org/TR/xml-infoset/
- Infoset describes "what all info can we get from a node (element, document, attribute...)"
- In other words: an application should not rely on any other info, such as attribute order etc.
- Any well-formed XML document conformant to XML Namespaces has its Infoset.

XML Infoset - structure

- *Infoset* comprises of *Information items*
- Infoset relates to document with expanded (resolved) entities
- We distinguish among infoset of document, element, attribut, character, PI, not-expanded entity, not-analysed entity, notation.

Canonical Form

- *Canonical XML Version 1.0, W3C Recommendation 15 March 2001*
- http://www.w3.org/TR/xml-c14n
- The goal of the Canonical Form is to describe criteria and algorithm how to define equivalence on XML documents that are "logically" the same and expose just differences in physical form (entities, attribute order, character encoding)
- Canonication "wipes-out" differences that are not significant for applications.
- Canonication is inevitable in some important applications such as information security, e.g. electronic signature of XML data (when calculating digest).

Canonical Form - principles

Main principles for constructing the canonical form of an XML document:

- encoding in UTF-8
- line breaks (CR, LF) normalized according to the algorithm mentioned in XML 1.0 Spec.
- attribute values normalized
- references to character and parsed entites replaced by their content
- CDATA section also replaced by their content
- prolog xml and DTD reference removed

Canonical Form - principles (contd)

- whitespaces outside of the root element normalized
- otherwise (except of line breaks), the whitespaces are preserved
- attribute values always in double quotes "
- special chars in attr. values replaced by refs to character entities
- superfluous NS declarations removed
- default attribute values added to all element where relevant
- attributes and NS declarations will be ordered lexikographically

Issues with Canonical Form

Certain information loss (mostly info from DTD):

- not-parsed entity (eg. binary ones) are not accessible anymore after canonicalization
- notations
- attribute types (incl. default values)