Processing of structured documents

Spring 2002, Part 2
Helena Ahonen-Myka

XML data modelling issues

- data model:
  - describes which information contained in XML documents is accessible
  - may be different for different processors and applications
    - format of input and output data
    - data that is used internally e.g. values of expressions

Data modelling issues

- the "normal" processing model:
  - an XML processor parses an XML document
    - checks at least the well-formedness
    - may also validate the document
    - provides the information of the document in some form to an application

XML 1.0 reporting requirements

- For instance:
  - an XML processor must always provide all characters in a document that are not part of markup to the application
  - a validating XML processor must inform the application which of the character data in a document is white space appearing within element content
  - an XML processor must normalize line-ends to LF before passing them to the application

XML 1.0 reporting requirements (ctnd.)

- A validating XML processor must include the replacement text of an entity in place of an entity reference
- an XML processor must supply the default value of attributes declared in the DTD for a given element type but not appearing in the element's start tag

XML data modelling issues

- several XML data models exist:
  - XML Information set (Infoset)
    - base for the other data models
    - describes information after parsing
  - PSVI (Post Schema Validation Infoset)
    - type information added
  - XQuery 1.0 and XPath 2.0 Data Model
    - also used in XSLT 2.0
    - input/output + internal representation
  - DOM

XML Information set

- W3C Recommendation: 24 Oct 2001
- purpose: to provide a set of definitions for use in other XML specifications that need to refer to the information in a well-formed XML document
- not meant to be exhaustive; not a set of minimum requirements that a processor has to return
- abstract definitions: no concrete interfaces etc. provided

7

XML Information set

- An XML document's **information set** consists of a number of information items
- an **information item** is an abstract description of some part of an XML document
  - mainly to be used in other specifications
- each information item has a set of associated named **properties**

8

XML Information set

- describes the tree structure provided by the processor (no special interface is specified)
  - e.g. entities expanded to their replacement text, attributes with their default values
- properties: e.g. for each element its child elements and attributes

9

Information items

- document information item
- element information items
- attribute information items
- processing instruction information items
- unexpanded entity reference information items
- character information items

10

Information items (cont.)

- comment information items
- document type declaration information item
- unparsed entity information items
- notation information items
- namespace information items

11

Document information item

- correponds to the document as a whole
  - do not confuse with the "real" root element (-> document element)
- there is exactly one document information item in the information set
- all information items are accessible from the properties of the document information item, either directly or indirectly through the properties of other information items

12

Document information item

- Properties:
  - children
  - document element
  - notations
  - unparsed entities
  - base URI
  - character encoding scheme
  - standalone
  - version
  - all declarations processed

13

Document information item

- children property: an ordered list of child information items, in document order:
  - exactly one element information item (=document element)
  - one processing instruction (PI) information item for each PI outside the document element (the same for comments)
    - comments and PIs within the DTD are excluded
  - a document type declaration information item (if there is a DTD)
- document element property:
  - The element information item corresponding to the document element

14

Element information items

- There is an element information item for each element appearing in the XML document
- one of the element information items is the value of the document element property of the document information item (root element)
- all other element information items are accessible recursively

15

Element information items

- An element information item has the following properties:
  - namespace name
  - local name
  - prefix
  - children (element, pi, character, comment)
  - attributes
  - namespace attributes
  - in-scope namespaces
  - base URI
  - parent

16

Attribute information items

- There is an attribute information item for each attribute (specified or defaulted) of each element in the document
  - including namespace declarations
- attributes declared in the DTD with no default value and not specified in the element's start tag are not represented by attribute information items

17

Attribute information items

- An attribute information item has the following properties:
  - namespace name, local name, prefix
  - normalized value
  - specified
    - Was the value specified in the start tag or defaulted from DTD?
  - attribute type (ID, IDREF, ENTITY, NMTOKEN,...)
  - references (target of IDREF = some ID)
  - owner element

18

Character information items

- there is a character information item for each data character that appears in the document
- each character is a logically separate information item
  - but XML applications are free to chunk characters into larger groups as necessary
- propertiers of a character information item:
  - character code
  - Is this character element content whitespace?
  - parent

19

Example

```
<?xml version="1.0"?>
<msg:message doc:date="19990421"
        xmlns:doc="http://doc.example.org/namespaces/doc"
        xmlns:msg="http://message.example.org/"
>Phone home!</msg:message>
```

20

The information items for the sample document

- A document information item
- an element information item with namespace name "http://message.example.org/", local part "message", and prefix "msg"
- an attribute information item with the namespace name "http://doc.example.org/namespaces/doc", local part "date", prefix "doc", and normalized value "19990421"

21

The information set for the sample document (cont.)

- three namespace information items for the http://www.w3.org/XML/1998/namespace, http://doc.example.org/namespaces/doc, http://message.example.org namespaces
- two attribute information items for the namespace attributes
- eleven character information items for the character data

22

What is not in the information set?

- For instance,
  - the document type name
  - the difference between the two forms of an empty element: <foo/> and <foo></foo>
  - the order of attributes within a start-tag
  - white space within start-tags (other than significant white space in attribute values) and end-tags
  - the difference between CR, CR-LF, and LF line termination

23

Data model of XPath

- defines
  - input for XPath, XSLT and XQuery
  - all values needed in the expressions of these specifications
- based on XML information set
  - augmented by possible schema validation information -> input = Post Schema-Validation Infoset (PSVI)

24

Data model of XPath

- data model supports also values that are not supported by XML information set, e.g.
  - well-formed document fragments, sequences of fragments, sequences of documents
  - atomic values (boolean, integer…), sequences of atomic values, sequences of mixing nodes and atomic values

25

Tree model

- A conceptual model: no particular implementation is assumed
- A tree that contains nodes (7 types):
  - document node
  - element nodes
  - attribute nodes
  - text nodes
  - namespace nodes
  - processing instruction nodes
  - comment nodes

26

For each node: tree properties and content

- "tree properties":
  - parent
  - children
  - attributes
  - namespaces
- the content of text, attribute, or element node can be interpreted in two ways:
  - as a string value: "123"
  - as a typed value: (integer) 123

27

Values of properties and content

- every value handled by the data model is a sequence of zero or more items
- an item is either a node or an atomic value
- a sequence cannot be a member of a sequence
- a single item appearing on its own is modeled as a sequence containing one item

28

Document order

- document order is defined on all the nodes in the document:
  - root node is the first node
  - element nodes in order of the occurrence of their start tags
  - attribute nodes and namespace nodes before the children of the element
  - namespace nodes before attribute nodes

29

Document node

- a tree whose root node is a document node is referred to as a *document*
  - otherwise the tree is a *fragment*
- the element node for the document element is a child of the document node
- other children:
  - processing instruction nodes
  - comment nodes
- string-value: concatenation of the string-values of all text node descendants of the document node in document order

30

Element nodes

- An element node for every element in the document
- children:
  - element nodes (subelements)
  - comment nodes
  - processing instruction nodes
  - text nodes (content)
- string-value:
  - concatenation of the string-values of all text node descendants of the element node in document order

31

Attribute nodes

- Each element node has an associated set of attribute nodes
  - the element node is the parent of each of these attribute nodes
  - but: an attribute node is not a child of its parent element
- a defaulted attribute is treated the same as a specified attribute

32

Attribute nodes

- if an attribute was declared for the element with the default #IMPLIED, but the attribute was not specified on the element, there is no attribute node for this attribute
- String-value: the normalized value as specified by the XML specification

33

Namespace nodes

- Each element has an associated set of namespace nodes
  - one for each distinct namespace prefix that is in scope for the element
  - one for the default namespace if one is in scope for the element
- The element is the parent of each of these namespace nodes, but a namespace node is not a child of its parent element
- string-value: the namespace URI

34

PI nodes, comment nodes

- There is a processing instruction node for every processing instruction
- there is a comment node for every comment
  - string-value: the content of the comment not including <!-- and -->
- … except for PIs and comments in document type declarations

35

Text nodes

- Character data is grouped into text nodes
- as much character data as possible is grouped into each text node
- string-value: the character data
- characters inside comments, processing instructions and attribute values do not produce text nodes

36

Example

- the specification contains an example (the same we had in the lecture):
  - http://www.w3.org/TR/query-datamodel/#d0e3694
  - See e.g. the string values of different nodes.
  - There's also a picture in the end.
  - Don't worry if you don't understand everything!

37

Data model issues

- we have seen...
  - general specification: XML Information Set
  - one specification that is based on the Information Set: XPath and XQuery data model

38