DOCUMENT SPEC

GRAPHS

VERTICES + EDGES (6 SPANNING TREE)

1. DOC SPEC SHOULD NOT INCLUDE AD HOC CRITERIA

2. LOCAL WAYS TO PROCESS DOCS SHOULD BE LARGELY INDEPENDENT OF PUBLIC SPEC OF A DOC

BUSINESS A

© North Carolina State University, All Rights Reserved
Uniform Resource Identifier

- URIs are abstract
- What matters is their (purported) uniqueness
- URIs have no proper syntax per se

Kinds of URIs
- URLs, as in browsing: not used in standards any more
- URNs, which leave the mapping of names to locations up in the air

Good design: the URI resource exists
- Ideally, as a description of the resource in RDDL
- Use a URL or URN
RDDL

Resource Directory Description Language

- Meant to solve the problem that a URI may not have any real content, but people expect to see some (human readable) content
- Captures namespace description for people
  - XML Schema
  - Text description
Module 3: XML Manipulation

Key XML query and manipulation languages include

- XPath
- XQuery
- XSLT
Metaphors for Handling XML: 1

How we conceptualize XML documents determines our approach for handling them

- **Text**: an XML document is text
  - Ignore any structure and perform simple pattern matches

- **Tags**: an XML document is text interspersed with tags
  - Treat each tag as an “event” during reading a document, as in SAX (Simple API for XML)
  - Construct regular expressions as in screen scraping
Metaphors for Handling XML: 2

- **Tree**: an XML document is a tree
  - Walk the tree using DOM (Document Object Model)
- **Template**: an XML document has regular structure
  - Let XPath, XSLT, XQuery do the work
- **Thought**: an XML document represents an information model
  - Access knowledge via RDF or OWL
Recursive Descent

/ Song / Song / Song

... Song

LG Song

< (en)  < (pa)  < (cpe)  < (him)

...
XPath

Used as part of XPointer, SQL/XML, XQuery, and XSLT

- Models XML documents as trees with nodes
  - Elements
  - Attributes
  - Text (PCDATA)
  - Comments
  - Root node: above root of document
Achtung!

- Parent in XPath is like parent as traditionally in computer science
- Child in XPath is confusing:
  - An attribute is not a child of its parent
  - Makes a difference for recursion (e.g., in XSLT apply-templates)
- Our terminology follows computer science:
  - e-children, a-children, t-children
  - Sets via et-, ta-, and so on

© Munindar P. Singh. CSC 513, Spring 2010
XPath Location Paths: 1

- Relative or absolute
- Reminiscent of file system paths, but much more subtle
  - Name of an element to walk down
    - Leading /: root
    - /: indicates walking down a tree
      - ..: currently matched (context) node
      - ...: parent node

© Munindar P. Singh. CSC 513, Spring 2010
XPath Location Paths: 2

- @attr: to check existence or access value of the given attribute
- text(): extract the text
- comment(): extract the comment
- [ ]: generalized array accessors
- Variety of axes, discussed below

例: `<Song lg='en'> SBR1 </Song>`
XPath Navigation

[] with numbers

- Select children according to position, e.g., [j], where j could be 1...last()

- Descendant-or-self operator, //
  - //elem finds all elems under the current node
  - //elem finds all elems in the document

- Wildcard, *:
  - collects e-children (subelements) of the node where it is applied, but omits the t-children
  - @*: finds all attribute values