

# Data-centric XML

## Information Set

# Overview

- Basis for use in other specifications (e.g. web services, Schema)
- Second edition February 2004
  - <http://www.w3.org/TR/xml-infoset/>
  - Covers XML 1.1 and namespaces 1.1
- “to refer to the information in a well-formed XML document”
  - not meant to be exhaustive; includes information with “expected usefulness in future specifications”
  - Validity is not a requirement, but namespace conformance is
- Infoset created through parsing, or other means
- Does not imply specific API or interfaces (e.g. tree API)
  - information might become available through events, queries, ...

# Structure

- Information consists of a set of **information items**
  - at least a **document information item**
- Each item has a set of associated **properties**
  - Property names are denoted in square brackets, **[thus]**
- Document must be namespace-conforming, and must not use relative namespace URIs
  - [namespace name] is the normalized value of the namespace attribute
- Entities may or may not be present
  - **unexpanded entity reference** information items used if needed
- Properties may have **unknown** or **no value**

# Document Information Item

- [children] contains exactly one element information item
- [document element]
- [notations]
- [unparsed entities]
- [base URI]
- [character encoding scheme]
- [standalone]
- [version]
- [all declarations processed]

# Element Information Item

- [namespace name]
  - no value if not qualified
- [local name]
- [prefix]
- [children]
  - list of information items, in document order (element, processing instruction, unexpanded entity reference, character, comment)
- [attributes]
  - unordered set of attribute information items
  - set might be empty
  - no namespace declarations
- [namespace attributes]
- [in scope namespaces]
  - unordered set of namespace information items
- [base URI]
- [parent]

# Attribute Information Item

- [namespace name], [local name], [prefix]
- [normalized value]
- [specified]
- [attribute type]
  - ID, IDREF, IDREFS, ENTITY, ENTITIES, NMTOKEN, NMTOKENS, NOTATION, CDATA, ENUMERATION
  - no value if there is no declaration
    - if [all declarations processed] is false, value is unknown
- [references]
  - unordered list if attribute is of type IDREF or IDREFS
  - no value or unknown otherwise
- [owner element]

# Processing Instruction Information Item

- [target]
- [content]
- [base URI]
- [notation]
- [parent]

# Unexpanded Entity Reference

- [name]
- [system identifier]
- [public identifier]
- [declaration base URI]
- [parent]



# Character Information Item

- [character code]
- [element content whitespace]
  - true if white-space character in element content
  - unknown if declarations have not been processed; mandatory for validating parsers
- [parent]

# Comment Information Item

- [content]
- [parent]

# Document Type Declaration

- [system identifier]
- [public identifier]
- [children]
  - An ordered list of processing instruction information items representing processing instructions appearing in the DTD, in the original document order. Items from the internal DTD subset appear before those in the external subset.
- [parent]

# Unparsed Entity

- [name], [system identifier], [public identifier]
- [declaration base URI]
- [notation name]
- [notation]

# Notation Information Items

- [name], [system identifier], [public identifier]
- [declaration base URI]

# Namespace Information Item

- [prefix]
- [namespace name]

# Conformance

- Specifications are conformant, not implementations
- Specifications must:
  - indicate which information items are needed to implement the specification
  - note any information needed from an XML document that is not in the infoset
  - note any difference in terminology compared to infoset terminology (this should be avoided)