

Grammars for XML

# XML Information Set

---

Lecture "XML in Communication Systems"  
Chapter 2

Dr.-Ing. Jesper Zedlitz  
Research Group for Communication Systems  
Dept. of Computer Science  
Christian-Albrechts-University in Kiel



# Recommended Reading

---

- T. Bray, J. Paoli, C.M. Sperberg-McQueen, E. Maler, F. Yergeau, J. Cowan (Eds.):  
*XML 1.1 (Second Edition), W3C Recommendation, 16 August 2006.*  
<http://www.w3.org/TR/xml11>
- J. Cowan, R. Tobin (Eds.):  
XML Information Set (Second Edition), W3C Recommendation, 4 February 2004.  
<http://www.w3.org/TR/xml-infoset>

# Overview

---

Informatik · CAU Kiel

1. Introduction
2. Information items
3. Information item names (Namespaces)
4. Further information item properties



Chapter 2.1

# Introduction

# Introduction

---

- What is XML Information Set?
  - A specification of **abstract data structures** describing the content of well-formed XML documents accessible to applications.
  - A specification describing the "output" of XML processors.
  - A W3C recommendation.

# Introduction

- From the XML specification:
  - "Each XML document has both a **logical** and a **physical** structure.
    - **Logically**, the document is composed of ..., **elements**, **comments**, ... and **processing instructions**, all of which are indicated in the document by explicit markup. (More "logical units" to come later; NL)
    - **Physically**, the document is composed of [storage] units called entities. ..."



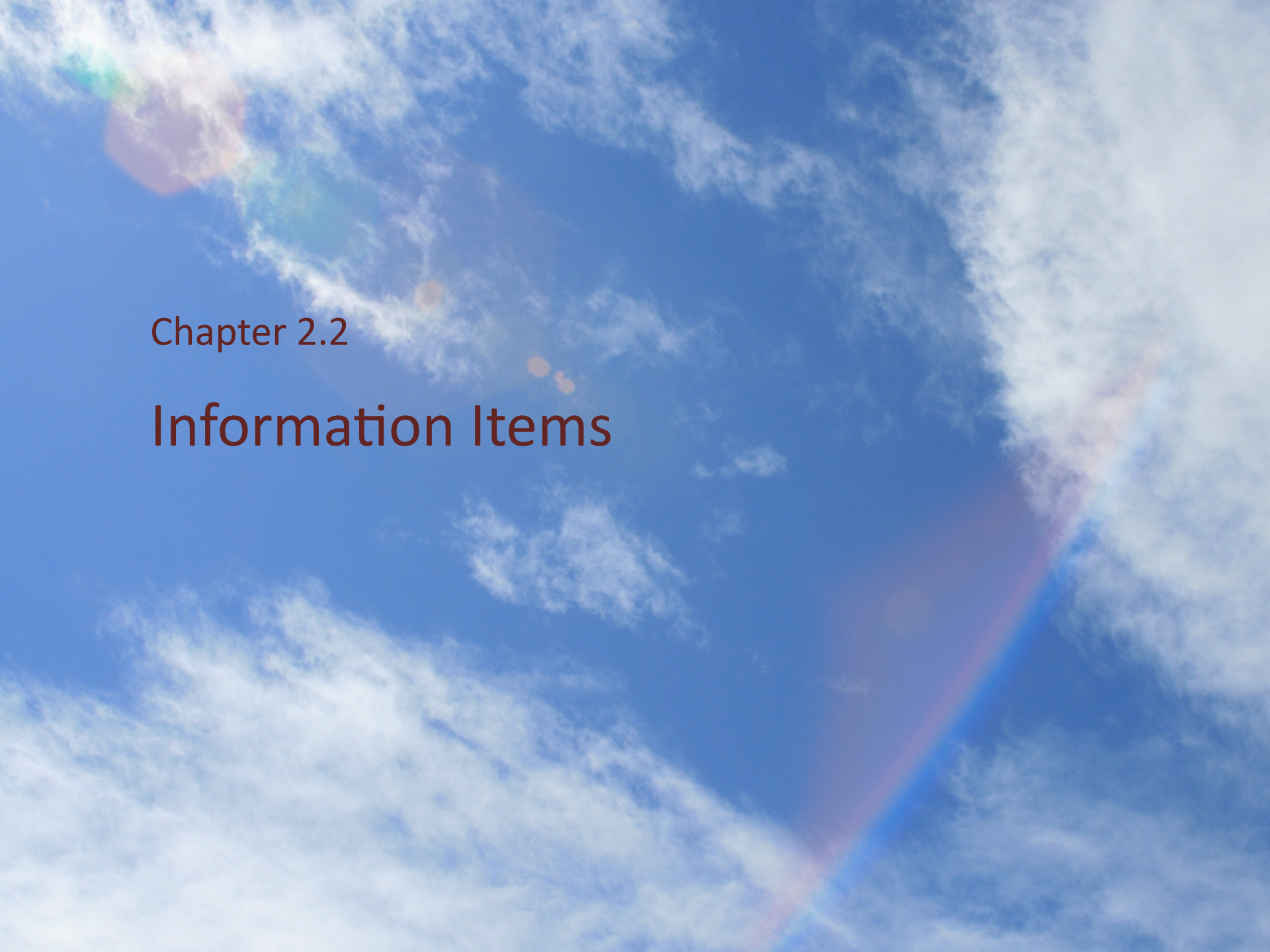
What is the **logical** structure of XML documents?

# Introduction

---

- What is an XML info set?
  - "An XML document's information set consists of a number of **information items**; the information set for any well-formed XML document will contain at least a **document information item** and several others. ... each information item has a set of associated named **properties**."



A vibrant blue sky with wispy white clouds. A prominent rainbow arches across the frame, starting from the bottom right and curving towards the top left. The colors of the rainbow are clearly visible, with red, orange, yellow, green, and blue bands. The overall scene is bright and cheerful.

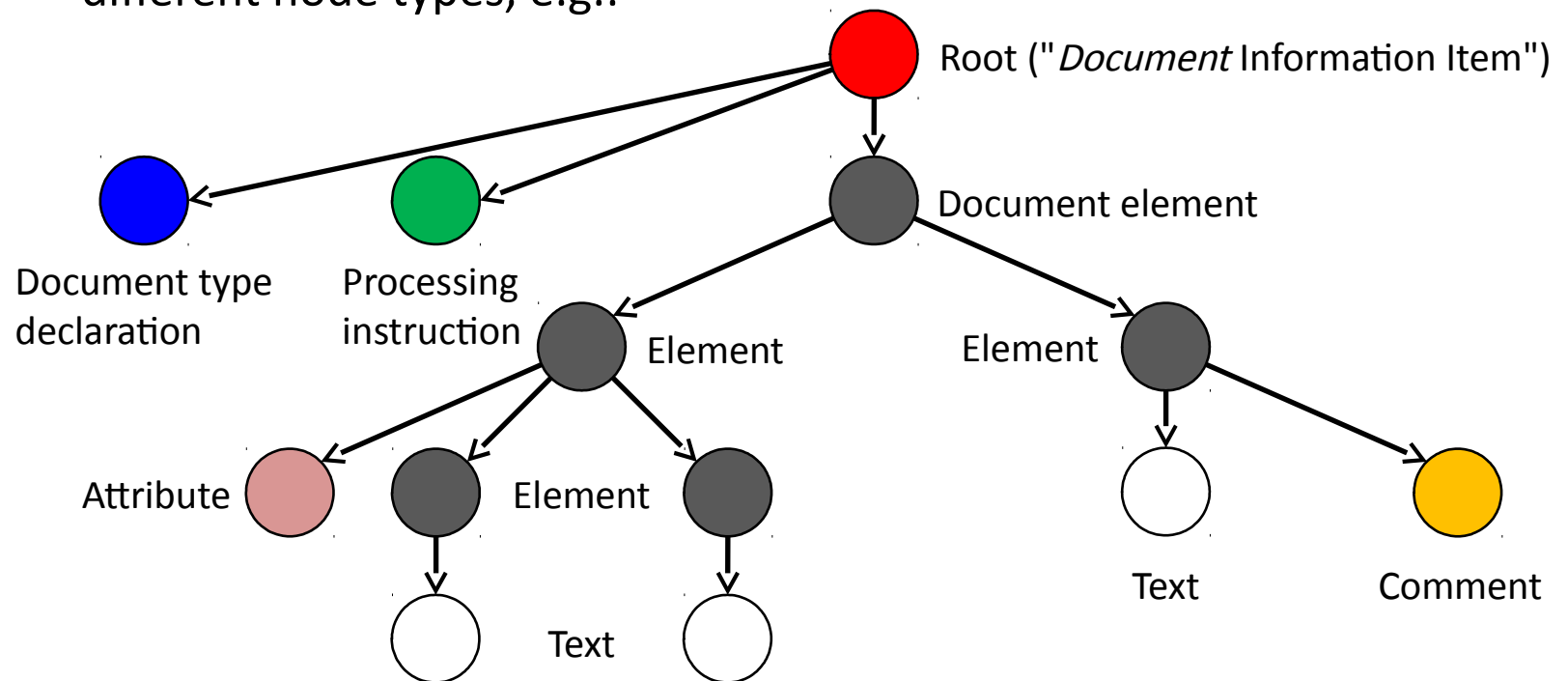
Chapter 2.2

# Information Items



# Information Items

- XML document
  - attributed, tree-like graph
  - different node types, e.g.:



# Information Items

- Types of **information items**
  - Document Information Item
  - Element Information Items
  - Attribute Information Items
  - Processing Instruction Information Items
  - Unexpanded Entity Reference Information Items
  - Character Information Items
  - Comment Information Items
  - Document Type Declaration Information Item
  - Unparsed Entity Information Items
  - Notation Information Items
  - Namespace Information Items

*We skip most of these!*

# Information Items

- Types of **information items**



## Document Information Item

### Element Information Item

- Attribute Information Item

- Processing Instruction

- Unexpanded Entity Reference

- Character Information Item

- Comment Information Items

- Document Type Declaration Information Item

- Unparsed Entity Information Items

- Notation Information Items

- Namespace Information Items

Is the root of the document: It has exactly one child *element information item*: the *document element information item*.

Precisely: The value of the [document element] property of the *document information item* is a single *element information item*. This is called the *document element information item*.

# Excursion: Terminology

- Inconsistent terminology between the DOM, XML, and XPath specs
  - "There is confusion between the terms **top level (document) element** (which is an element node) with **root** of the document (which is not)."
  - Clarification:
    - The **root** represents the document itself.
    - The **document element** is the element that has all other elements within the document as descendants.
    - The **root** can have other children besides the document element, e.g. comments and processing instructions.

# Information Items

- Types of **information items**

- Document Information Item



- Element Information Items

- Attribute Information Item

- Processing Instruction

- Unexpanded Entity Reference Information Items

- Character Information Items

- Comment Information Items

- Document Type Declaration Information Item

- Unparsed Entity Information Items

- Notation Information Items

- Namespace Information Items

There is an *element information item* for each element appearing in the XML document.



# Information Items

- Types of **information items**

- Document Information Item
- Element Information Items



## Attribute Information Items

### Processing Instruction Information Items

- Unexpanded Entity Reference
- Character Information Items
- Comment Information Items
- Document Type Declaration
- Unparsed Entity Information Items
- Notation Information Items
- Namespace Information Items

There is an *attribute information item* for each attribute (specified or defaulted) of each element in the document, including those which are namespace declarations. The latter however appear as members of an element's [namespace attributes] property rather than its [attributes] property.

# Information Items

- Types of **information items**

- Document Information Item
- Element Information Items
- Attribute Information Items
- Processing Instruction Information Item
- Unexpanded Entity Reference
- Character Information Items
- Comment Information Item
- Document Type Declaration Information Item
- Unparsed Entity Information Items
- Notation Information Items
- Namespace 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 or desirable. (In this lecture, we prefer a *text information item*.)

# Information Items

- Types of **information items**

- Document Information Item
- Element Information Items
- Attribute Information Items
- Processing Instruction Information Items
- Unexpanded Entity Reference Information Items
- Character Information Items
- Comment Information Items
- Document Type Declaration Information Item
- Unparsed Entity Information Item
- Notation Information Item
- Namespace Information Items

Each element in the document has a namespace information item for each namespace that is in scope for that element.





An aerial photograph of a large, irregularly shaped ice floe in the ocean. The ice floe is a light, mottled greenish-grey color, contrasting with the darker, choppy water surrounding it. The floe has a complex, organic shape with several smaller, detached pieces nearby. The water is a deep, dark teal color with visible ripples and small waves.

Chapter 2.3

# Information Item Names (Namespaces)

# Information Item Names

---

- Remember:
  - "An XML document's information set consists of a number of **information items**; the information set for any well-formed XML document will contain at least a document information item and several others. ... each information item has a set of associated named **properties**."
- Three important name-related properties
  - local name
  - expanded name
  - namespace URI



# Information Item Names

---

Informatik · CAU Kiel

- Problem
  - How to provide unique names, when "mixing" XML documents?

```
<?xml version="1.0" encoding="UTF-8"?>
<Book>
  <ISBN>0743204794</ISBN>
  <author>Kevin Davies</author>
  <title>Cracking the Genome</title>
  <price>20.00</price>
</Book>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<html>
  <head>
    <title>My home page</title>
  </head>
  <body>
    <p>My hobby</p>
    <p>My books</p>
  </body>
</html>
```

"local  
names"

```
<?xml version="1.0" encoding="UTF-8"?>
<html>
  <head>
    <title>My home page</title>
  </head>
  <body>
    <p>My hobby</p>
    <p>My books
      <Book>
        <ISBN>0743204794</ISBN>
        <author>Kevin Davies</author>
        <title>Cracking the Genome</title>
        <price>20.00</price>
      </Book>
    </p>
  </body>
</html>
```

Remark: This is called "mixed content"—  
text content and element content

Both XML validator and application program  
need context information to distinguish between  
*HTML page title* and *book title*!

# Namespaces

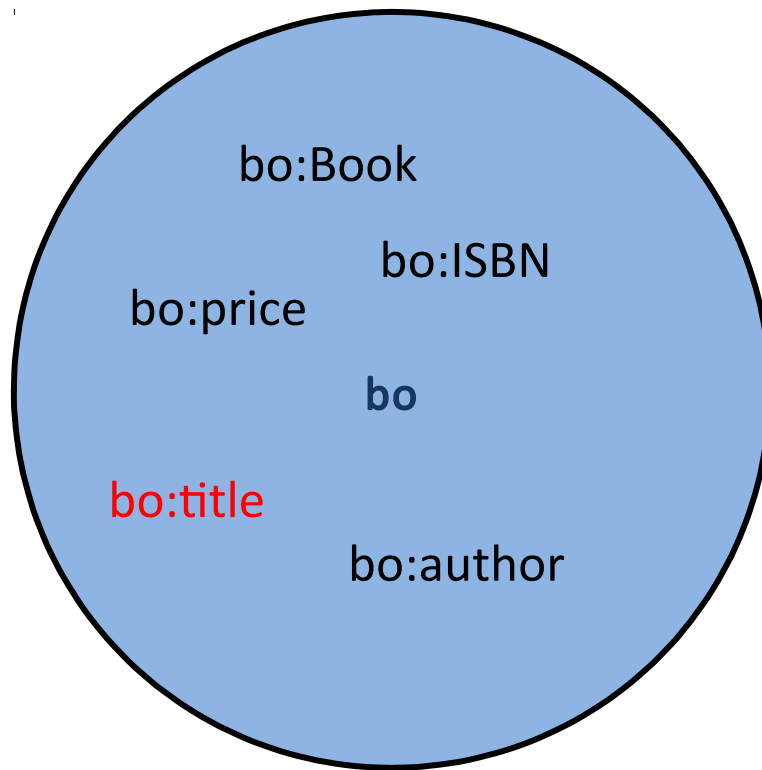
- How the web works
  - Individually created documents
  - Distributed creation of knowledge and lazy integration
  - Problem: Vocabulary collisions!
- Two-step solution
  1. Expand local names by locally unique name prefixes
  2. Bind local prefixes to globally unique URIs

```
<?xml version="1.0" encoding="UTF-8"?>
<xhtml:html>
  <xhtml:head>
    <xhtml:title>My home page</xhtml:title>
  </xhtml:head>
  <xhtml:body>
    <xhtml:p>My hobby</xhtml:p>
    <xhtml:p>My books
      <bo:Book>
        <bo:ISBN>0743204794</bo:ISBN>
        <bo:author>Kevin Davies</bo:author>
        <bo:title>Cracking the Genome</bo:title>
        <bo:price>20.00</bo:price>
      </bo:Book>
    </xhtml:p>
  </xhtml:body>
</xhtml:html>
```

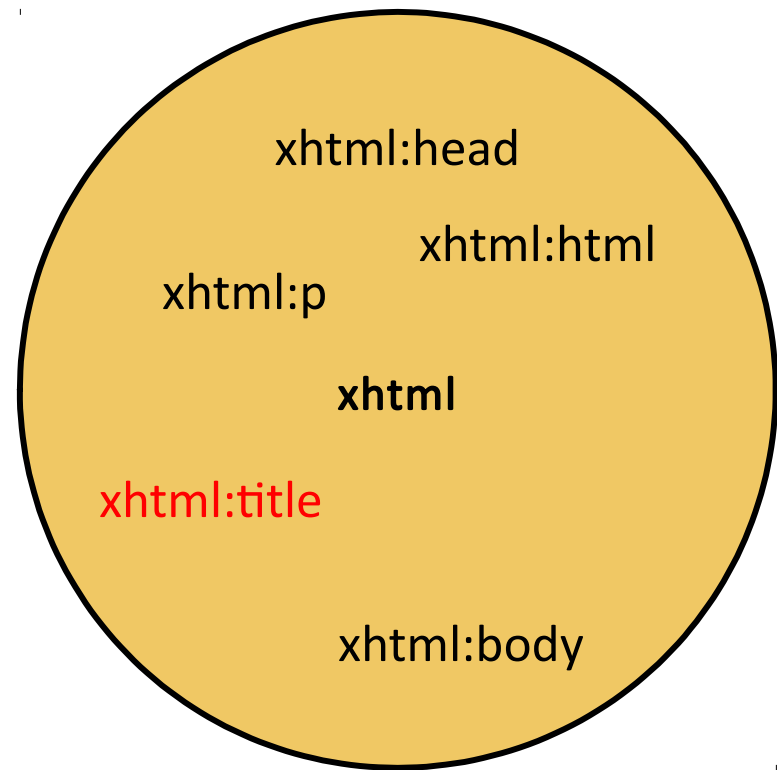
locally unique  
"expanded names"



# Namespaces



vocabulary bo



vocabulary xhtml

- But who guarantees uniqueness of prefixes?

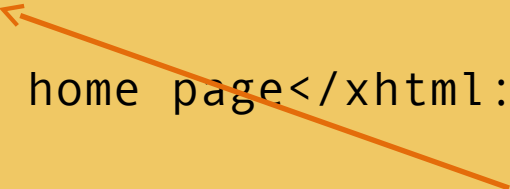
# Namespaces

- Give prefixes only local relevance in an instance document
  - Associate local prefix with global namespace name:  
a unique name for a namespace
  - Uniqueness is guaranteed by using a URI (preferably URN)  
in domain of the party creating the namespace.
  - URI doesn't have any meaning,  
i.e. doesn't have to resolve into anything.



An XML namespace is a collection of names, identified by a URI reference, which are used in XML documents as element and attribute names.

```
<?xml version="1.0" encoding="UTF-8"?>
<xhtml:html
  xmlns:xhtml="http://www.w3c.org/1999/xhtml"
  xmlns:bo="http://www.nogood.com/Book">
  <xhtml:head>
    <xhtml:title>My home page</xhtml:title>
  </xhtml:head>
  <xhtml:body>
    <xhtml:p>My hobby</xhtml:p>
    <xhtml:p>My books
      <bo:Book>
        <bo:ISBN>0743204794</bo:ISBN>
        <bo:author>Kevin Davies</bo:author>
        <bo:title>Cracking the Genome</bo:title>
        <bo:price>20.00</bo:price>
      </bo:Book>
    </xhtml:p>
  </xhtml:body>
</xhtml:html>
```



namespaces

```
<?xml version="1.0" encoding="UTF-8"?>
<xhtml:html
  xmlns:xhtml="http://www.w3c.org/1999/xhtml">
  <xhtml:head>
    <xhtml:title>My home page</xhtml:title>
  </xhtml:head>
  <xhtml:body>
    <xhtml:p>My hobby</xhtml:p>
    <xhtml:p>My books
      <bo:Book
        xmlns:bo="http://www.nogood.com/Book">
          <bo:ISBN>0743204794</bo:ISBN>
          <bo:author>Kevin Davies</bo:author>
          <bo:title>Cracking the Genome</bo:title>
          <bo:price>20.00</bo:price>
        </bo:Book>
      </xhtml:p>
    </xhtml:body>
  </xhtml:html>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<xhtml:html
  xmlns:xhtml="http://www.w3c.org/1999/xhtml">
  <xhtml:head>
    <xhtml:title>My home page</xhtml:title>
  </xhtml:head>
  <xhtml:body>
    <xhtml:p>My hobby</xhtml:p>
    <xhtml:p>My books
      <Book
        xmlns="http://www.nogood.com/Book">
          <ISBN>0743204794</ISBN>
          <author>Kevin Davies</author>
          <title>Cracking the Genome</title>
          <price>20.00</price>
        </Book>
      </xhtml:p>
    </xhtml:body>
  </xhtml:html>
```



```
<?xml version="1.0" encoding="UTF-8"?>
<html
  xmlns="http://www.w3c.org/1999/xhtml">
  <head>
    <title>My home page</title>
  </head>
  <body>
    <p>My hobby<p>
    <p>My books
      <Book
        xmlns="http://www.nogood.com/Book">
          <ISBN>0743204794</ISBN>
          <author>Kevin Davies</author>
          <title>Cracking the Genome</title>
          <price>20.00</price>
        </Book>
      </p>
    </body>
  </html>
```

# Namespaces

- What do namespace URI's point to?
  - The "abstraction" camp
    - A namespace URI is the id for a concept, it shouldn't resolve to anything
  - The "orthodox" camp
    - It should resolve to a schema (xml schema)
  - The "liberal" camp
    - It should resolve to many things



Chapter 2.4

# Further Information Item Properties

# Information Item Properties

- Properties depend on type of information item
  - e.g. element information item
    - [namespace name], [local name], [prefix]
    - [children] An ordered list of child information items, in document order. This list contains element, processing instruction, unexpanded entity reference, character, and comment information items ...
    - [attributes] An unordered set of attribute information items, one for each of the attributes ...
    - [namespace attributes] An unordered set of attribute information items, one for each of the namespace declarations
    - [base URI] The base URI of the element.
    - [parent] The document or element information item which contains this information item in its [children] property.

# Information Item Properties

- Properties depend on type of information item
  - e.g. attribute information item
    - [namespace name], [local name], [prefix]
    - [normalized value] The normalized attribute value
    - [specified] A flag indicating whether this attribute was actually specified in the start-tag of its element, or was defaulted
    - [attribute type] An indication of the type declared for this attribute in the DTD. Legitimate values are ID, IDREF, IDREFS, ENTITY, ENTITIES, NMTOKEN, NMTOKENS, NOTATION, CDATA, and ENUMERATION. ... Applications should treat no value and unknown as equivalent to a value of CDATA. The value of this property is not affected by the validity of the attribute value.
    - [references] For attributes typed IDREF, IDREFS, ENTITY, ENTITIES, or NOTATION
    - [owner element] The element information item which contains this information item in its [attributes] property.