XML & XHTML:
An Overview

R. Scott Granneman

Senior Consultant, Internet Services

Bryan Consulting • www.bryanconsulting.com

HTML

- HyperText Markup Language
- Standard publishing language of the Web
- 4.01 is the latest spec ... & the *last* spec for HTML, ever

CSS

- Cascading Style Sheets
- Goal is to separate data from presentation
- Use a single style sheet to apply to all pages on a Web site
  - o Formatting
  - o Positioning

DOM

- Document Object Model
- Allows developer to ...
  - Name any object on a Web page
  - Write scripts that interact with & affect any named object
- In other words, it allows developers to exert programatic control over documents
- For more info, see: www.w3.org/TR/WD-DOM/introduction.html

XML

- eXtensible Markup Language
- HTML describes content on a Web page in terms of display and interaction:
  - o \<P> or \<TABLE> or \<FORM>
- XML describes content in terms of what data is being described:
  - o \<AUTHOR> or \<GPA> or \<FIRSTNAME>

HTML Compared to XML
(The Amazon Example)

HTML

```
<p>
  R. Scott Granneman
  <br>HTML How-To
  <br>$10
</p>
```

XML

```
<author>
  R. Scott Granneman
</author>
<title>
  HTML How-To
</title>
<price>
  $10
</price>
```

Sample XML

```xml
<addressbook>
  <entry>
    <name>Bill Gates</name>
    <email>billg@microsoft.com</email>
  </entry>
  <entry>
    <name>Jerry Bryan</name>
    <email>jerry@bryanconsulting.com</email>
  </entry>
</addressbook>
```

HTML Works With Other Technologies

- HTML describes data
- CSS describes presentation of that data
- Scripts describe the handling or behavior of that data, as prescribed by the DOM

XML Works With Other Technologies

- XML describes data
- CSS describes presentation of that data
- Scripts describe the handling or behavior of that data, as prescribed by the DOM

XML is "Extensible"

- XML is extensible because you – or your industry – can create the markup tags you need

- Another bookstore example

SportsML

- "SportsML aims to be the global XML standard for the interchange of sports data. Designed to be as easy to understand and implement as possible, SportsML allows for the exchange of sports scores, schedules, standings, and statistics for a wide variety of competitions." (www.sportsml.com)

SportsML Example

```
<team>
  <team-metadata team-key="l.nba.com-t.10" team-source="iptc.org">
   <name first="Chicago" last="Bulls"/>
  </team-metadata>
  <team-stats standing-points=".111" games-back="5.5" streak="L-6">
   <outcome-totals wins="1" losses="8"/>
   <outcome-totals duration-scope="events-overtime"/>
   <outcome-totals competition-scope="division"/>
   <outcome-totals/>
   <outcome-totals duration-scope="events-most-recent-10" wins="1"
losses="8"/>
   <outcome-totals alignment-scope="events-home" wins="1" losses="3"/>
   <outcome-totals alignment-scope="events-home" duration-
scope="events-overtime"/>
   <rank value="8"/>
  </team-stats>
</team>
```

XML is Flexible

- XML files can be ...
  - o Displayed to humans via Web pages
  - o Processed as data by programs
    - AKA
      - ➤ The "browserless Web" or
      - ➤ "3rd generation Internet" or
      - ➤ "Semantic Web"
    - Another bookstore example

What is XML Good For?

- Working with structured data
- Interchanging data
- What are some examples?

A Note on the Longevity of HTML

- XML is not going to "kill HTML"
- HTML is ideal for many situations
- When should you use XML? HTML?
  - o Family pictures
  - o Contact list
  - o Small online store selling dolls
  - o Essays & opinions

XHTML

- What it is
- Why use it
- Flexibility of XHTML
- DTDs
- HTML to XHTML
- XHTML 1.1
- XHTML 2.0
- Tools

XHTML

- eXtensible HyperText Markup Language
- Reformulation of HTML 4.01 as a subset of XML
- If you know HTML 4.01 & CSS, you shouldn't have any problems
- XHTML 1.0 released on 26 January 2000 as a recommendation by W3C

HTML * XML = XHTML

- XHTML is a bridge from HTML to XML
  - HTML is easy & known
  - XML is powerful
- Documents in XHTML are both ...
  1. readable by older browsers AND
  2. usable as we move ahead ...
- ... thus maintaining backward & forward compatibility

So Why Use XHTML?

HTML is a Mess

- HTML was originally designed to exchange narrative documents between scientists

- HTML has been forced into shapes that were never considered by its creators, esp. concerning the exchange of data

- Sloppy & incorrect markup makes things difficult for browsers, users, & developers

XHTML is the Future

- HTML is now at version 4.01
- There will be no further revisions
- W3C is focusing all future revisions on XML & one of its subsets, XHTML

X is for "eXstensible"

- XHTML, like XML, can be extended by users
  - o XHTML starts with a set of known HTML elements & attributes
  - o New tags & attributes can be defined & added to those that already exist
  - o Developers can mix known HTML 4.0 elements with elements from other XML languages

Easier Support of Wireless Devices

- Currently, HTML-rendering engines have to allow for sloppy & incorrect markup
- Thus, they have to be a certain size
  - Gecko is less than 2 MB
  - mshtml.dll is 2.6 MB
- Since XHTML requires conformance, the rendering engines can be smaller
- Thus, they can more easily fit into smaller devices, especially those that are wireless

What is Modularization?

- Modularization breaks XHTML up into smaller element sets, each with a specific type of functionality
- For more information, see www.xmlhack.com/read.php?item=118

Some Modules

- Structure (body, head, html, title)
- Image (img)
- Hypertext (a)
- List (ol, ul, li)
- Forms
- Table
- Intrinsic Events (onXXX)

- Metainformation (meta)
- Scripting (script)
- Stylesheet (style)
- Link (link)
- Object
- Bi-directional Text
- Text (p, br, code, h1)

Why Do We Need Modularization?

- As XHTML moves from browsers to other platforms, not all XHTML elements will be required

- For example, each of the following may only support a subset of XHTML
  - Palm Pilot
  - Cell phone
  - WebTV

- Now, different devices using the *tables* module will all use the same tables definition

Basic XHTML Document Structure

```
<!DOCTYPE ...>
<html ...>
    <head>...</head>
    <body>...</body>
</html>
```

What is a DTD?

- Document Type Definition
- Describes the allowed legal
  - structure,
  - elements, &
  - attributes
- of a markup language (HTML, XHTML, etc.) used in a document that complies to the DTD

HTML DTDs

- Transitional:
  http://www.w3.org/TR/REC-html40/loose.dtd

- Strict:
  http://www.w3.org/TR/REC-html40/strict.dtd

- Frameset:
  http://www.w3.org/TR/REC-html40/frameset.dtd

HTML Transitional DTD Examples

- Defines required structure:
  o <html> at beginning & </html> at end
- Defines allowed elements:
  o <hr>
- Defines allowed attributes (& values) for each element:
  o align (left|center|right)
  o size %Pixels;
  o width %Length;

DTD's Purpose = Interoperability

- HTML DTDs are standardized for everyone, regardless of industry
- XML & XHTML DTDs are different
  - Industries must agree on a shared DTD
  - Once a DTD is agreed upon, companies can share data & information efficiently

XHTML DTDs

- DTD must be declared in 2nd line of document
- XHMTL 1.0 specifies 3 XHTML DTDs:
  - o Strict
  - o Transitional
  - o Frameset

DTD: XHTML 1.0 Strict

- Denotes
  - Clean markup
  - CSS instead of descriptive markup
  - No use of deprecated tags (like <CENTER>) or attributes

```
<!DOCTYPE html PUBLIC
    "-//W3C/DTD XHTML 1.0 Strict//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
```

Strict Requires Use of Namespace

- Must include the XHTML namespace in the <html> tag

- Why a namespace?
  - Browsers used to "extend" HTML willy-nilly
  - Namespace tells browser where elements & attributes used in the doc come from

```
<html
    xmlns="http://www.w3.org/1999/xhtml"
    xml:lang="en" lang="en">
```

DTD: XHTML 1.0 Transitional

- Denotes
  - o Descriptive markup + little or no CSS
  - o Use of deprecated tags & attributes

```
<!DOCTYPE html PUBLIC
    "-//W3C//DTD XHTML 1.0 Transitional//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
```

DTD: XHTML 1.0 Frameset

- Denotes frameset

```
<!DOCTYPE html PUBLIC
    "-//W3C//DTD XHTML 1.0 Frameset//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-frameset.dtd">
```

HTML to XHTML in 7 Easy Steps

Lowercase Elements & Attributes

- XML & therefore XHTML is case-sensitive
- Elements & attributes must be lowercase
- Values may be any case necessary

- HTML $\rightarrow$ XHTML
  - <TD BGCOLOR="#FFCC33"> $\rightarrow$ <td bgcolor="#FFCC33">
  - <IMG SRC="me.JPG"> $\rightarrow$ <img src="me.JPG">

Nest Elements Without Overlapping

- Browsers currently allow overlapping elements

- XHTML forces nesting: last opened is the first closed

- HTML $\rightarrow$ XHTML
  - o `<p>Be <b>good!</p></b>` $\rightarrow$
    `<p>Be <b>good!</b></p>`

Close All Non-Empty Elements

- Browsers currently allow HTML developers to skip some closing elements
- XHTML requires that all elements must be closed

- HTML $\rightarrow$ XHTML
  - o <p>Lorem ipsum $\rightarrow$ <p>Lorem ipsum</p>

Non-Empty Elements Requiring Closure

- <basefont>
- <body>
- <dd>
- <dt>
- <head>
- <html>

-
- <p>
- <td>
- <th>
- <tr>

- All require a closing tag!

Terminate Empty Elements

- Empty elements lack a closing tag & do not contain content
  - o <p> is not empty ... it contains text
  - o <br> is empty - no content is "in" it
  - o Notice the space before the "/" !

- HTML $\rightarrow$ XHTML
  - o
 $\rightarrow$
  - o <img src="pic.gif"> $\rightarrow$

Empty Elements Requiring Termination

- <area>
-
- <frame>
- <hr>
- <img>

- <input>
- <link>
- <meta>
- <option>

- All require a " /" before the closing ">" (notice the space!)

Quote All Attribute Values

- Quote every value, even if it's numeric

- HTML → XHTML
  - ◦ <img src=pic.gif border=0> →

Cannot Minimize Attribute-Value Pairs

- An attribute is minimized when there is only one possible value

- HTML currently allows stand-alone attributes

- XHTML requires that developers expand stand-alone attributes

- HTML → XHTML
  - o <td nowrap> → <td nowrap="nowrap">
  - o <input ... checked> → <input ... checked="checked">

Don't Nest Certain Elements

- <a> cannot contain another <a>
- <pre> cannot contain <img>, <object>, <big>, <small>, <sub>, or <sup>
- <form> cannot contain another <form>
- Inline elements cannot contain block-level elements
  - Inline elements: <a>
 <frame> <hr> <img> <input> <link> <meta> <option> <span>

Sample XHTML: Hello World!

```
<?xml version="1.0"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
<head>
    <title>Hello World!</title>
</head>
<body>
    <p>Hello World!</p>
    <img src="http://www.w3.org/Icons/valid-xhtml10" />
</body>
</html>
```

What's the First Line For?

- What's this?: <?xml version="1.0"?>
- All XML documents must begin with that line
- XHTML are XML docs, ergo ...
- Problem: older browsers (version 4 & below) freak when they see that line at the top of the page
- Solutions:
  - Ignore older browsers
  - Leave off the line & use XHTML 1.0 transitional
  - Two sites

Hello World! Analysis

- Notice:
  - DTD (Using strict instead of transitional)
  - Declaration of XHTML namespace in <html> element
  - Lowercase tags & attributes (Notice that uppercase is allowed in attribute value)
  - Closure of <img> tag
  - How similar it is to HTML!

Browser Support for XHTML

- Give your XHTML document a ".html" or ".htm" extension
  - o Your browser will display the page
  - o Your browser will ignore what it does not understand

XHTML 1.1

Changes from XHTML 1.0

- As of 6 April 2001, XHTML 1.1 is a W3C Proposed Recommendation

- XHTML 1.1 is a refinement of XHTML 1.0 Strict
  - What was deprecated in HTML 4.0 & XHTML 1.0 is now removed in XHTML 1.1
  - You must use CSS for presentation

Elements Removed From XHTML 1.1

- <basefont>
- <center>
- <font>
- <frame>
- <frameset>

- <iframe>
- <noframes>
- <object>
- <strike>
- <u>

Attributes Removed From XHTML 1.1

- <a> -target
- <body> -background, -bgcolor
- <div> -align
- <hr> -align, -noshade, -size, -width
- <img> -align, -border, -hspace, -vspace
- -type, -value
- <p> -align
- <table> -align, -bgcolor
- <td> -bgcolor, -height, -nowrap, -width

XHTML 2.0

Changes From XHMTL 1.x

- XHTML 2.0 spec introduced in August 2002
- 6th revision came out in July 2004
- Goal:
  - Introduce new, needed features while migrating away from deprecated features of HTML 4 and XHTML 1.0
  - Normalize XHTML so that things are more logical
  - **Not** intended to be backward compatible with existing Web pages

Changes in Elements & Attributes

- Deprecated HTML elements & attributes are gone for good

- "applet" & "img" elements are gone
  - All images are equivalent to content, so you can put the "src" attribute on any element
  - If image is available, use it; otherwise, use element's content
    - <p src="map.gif">Exit & turn left to 123 Main St.</p>

- "href=" attribute works with every element, so anything can be a link

Movements Towards XML

- Forms are replaced with XForms (www.w3.org/TR/xforms)
- Frames (not in 1.1) implemented via XFrames (www.w3.org/TR/XFrames)
- Javascript events are replaced with XML Events (www.w3.org/TR/xml-events)

Tools

Validate Your XHTML (& HTML)

- W3C's Validator
  - ○ Free service from the W3C
  - ○ Check your document's validity
  - ○ Point to a URL or upload a document
  - ○ Lists errors & problems
- http://validator.w3c.org

Validate Your CSS

- W3C's CSS Validator
  - o Free service from the W3C
  - o Check your CSS document's validity
  - o Point to a URL or upload a document
  - o Lists errors & warnings
- http://jigsaw.w3c.org/css-validator

Convert HTML to XHTML with Tidy

- Features
  - Clean up code
  - Convert HTML to XHTML
  - Use a configuration file to organize your settings
  - Open Source, free, & multi-platform!
- Software & info
  - http://www.w3c.org/People/Raggett/tidy/
  - http://tidy.sourceforge.net
  - http://www.granneman.com/webdev/editors/tidy/

Further Information

- www.granneman.com/usefullinks/development/coding/xhtml/ & www.granneman.com/usefullinks/development/coding/xml/
- Specifications: www.w3.org
- Tutorials: www.w3schools.com & www.irt.org
- XHTML listserv: www.yahoogroups.com/group/XHTML-LXML
- Information: www.xml.com & www.builder.com/Authoring/XmlSpot/
- Listservs: lists.w3.org/Archives/Public/

Books: Basic

- *HTML 4 For The World Wide Web, 4th Edition: Visual QuickStart Guide* by Elizabeth Castro

- *DHTML And CSS For The World Wide Web: Visual QuickStart Guide* by Nancy Davis

- *XML For The World Wide Web: Visual QuickStart Guide* by Elizabeth Castro

Books: Advanced

- *HTML & XHTML: The Definitive Guide* by Chuck Musciano & Bill Kennedy
- *XHTML: Moving Toward XML* by Simon St. Laurent
- *Cascading Style Sheets: The Definitive Guide* by Eric A. Meyer
- *The XML Bible* by Elliotte Rusty Harold

Thank You!

- Final thought: XML & XHTML offer exciting opportunities for businesses & developers

- Email me: scott@granneman.com

- Visit my Web site: www.granneman.com

- Join GranneNotes, my listserv for people interested in technology & the Internet, by visiting www.granneman.com & signing up

Licensing of this work

This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/1.0 or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

In addition to the rights and restrictions common to all Creative Commons licenses, the Attribution-ShareAlike License features the following key conditions:

Attribution. The licensor permits others to copy, distribute, display, and perform the work. In return, licensees must give the original author credit.

Share Alike. The licensor permits others to distribute derivative works under a license identical to the one that governs the licensor's work.

Questions? Email scott@granneman.com