DOCUMENT FORMATS FOR THE WEB

V Quint

Rapporteur: Dr Paul Watson

Document Formats for the Web

Vincent Quint

INRIA / W3C

Outline

1. HyperText Markup Language: **HTML**
2. Cascading Style Sheets: **CSS**
3. Extensible Markup Language: **XML**
4. Mathematical markup Language: **MathML**
5. Document Object Model: **DOM**

Original Web technology

- Addressing: **URL**
- Communication: **HTTP**
- Document format: **HTML**

Original HTML

Created by Tim Berners-Lee at CERN

- Representation of word processing documents
- Simple links (go to)

Syntax derived from SGML

Images added later

Versions of HTML

HTML 2.0 (13 Oct 94)

- Published by IETF (RFC 1866)
- Specified by an SGML DTD
- Specify the commonly used HTML tags in mid-94

HTML 3.2 (14 Jan 97)

- Published by W3C as a Recommendation
- Formalization of current practice
- Tables, applets, text flowing around images

HTML 4.0 (first draft 2 Apr 97)

HTML: general principles

General principles:

- Structure and presentation should be separated
- Universal accessibility to the Web:
  - People with disabilities
  - Internationalization

- HTML should remain simple to use

HTML 4.0

Extensions to HTML 3.2

- Multimedia objects
- Scripting
- Rich forms and interactive documents
- Richer tables
- Frames and subsidiary windows
- Support for style sheets
- Better access to HTML for people with disabilities
- Support for internationalization

Cascading Style Sheets

Style sheets describe how documents are presented

- on screens,
- in print,
- or how they are pronounced.

Cascade: more than one style sheet can apply to a document, from different sources

- document designer
- browser
- reader

CSS basics

CSS is a simple, declarative style sheet language

```
selector { property: value}
```

A selector identifies the elements to which a rule apply:

- Element type,
  Attribute (CLASS, ID),
  Structural context,
  Pseudo–class (link, visited, active),
  Pseudo–element (initial)

Style can be: external, embedded, interleaved

CSS1

Properties in CSS1:

- fonts, colors, backgrounds, margins, borders, padding, floating, block, inline, list–style

Examples:

```
H1 { color: red }
IMG.logo { float: right }
#xyz { font-variant: small-caps }
H1 STRONG { font-style: italic }
A:visited { border: solid red }
P:first-line { font-variant: small-caps }
```

CSS extensions

Printing: page breaks, page boxes, media dependent style sheets, alternate print document

Positioning: positioning and visibility of HTML elements in 2.5–dimensional space

ACSS: Aural Cascading Style Sheets

Fonts: Web fonts through CSS

Benefits from CSS

Style sheets allow to separate style from content

Style sheets improve

- Presentation
- Maintainability
- Accessibility
- Performance

Extensible Markup Language (XML)

- A data format for structured document interchange on the Web
- Developed for applications that require functionality beyond the current HTML
- XML is a clean subset of SGML specially designed for Web applications
- Can be parsed without prior knowledge

XML vs. HTML

XML differs from HTML in three major respects:

- Information providers can define new tag and attribute names at will.
- Document structures can be nested to any level of complexity
- Any XML document can contain an optional description of its grammar (DTD) for use by applications that need to perform structural validation

XML-link: Advanced linking

- Location–independent naming
- Bidirectional links
- Links that can be specified and managed outside of documents to which they apply
- N–ary hyperlinks (e.g., rings, multiple windows)
- Aggregate links (multiple sources)
- Transclusion (the link target document appears to be part of the link source document)
- Attributes on links (link types)

Mathematical Markup Language (MathML)

Goal of MathML:

- enable mathematics to be served, received, and processed on the Web

MathML

- an XML application
- describes mathematical notation
- captures both structure and content

MathML: presentation and content

Presentation tags:

25 tags describe abstract notational structures

MI, MO, MN, MFRAC, MROOT, MROW, MSUBSUP, MUNDEROVER

Content tags:

50 tags unambiguously specify the intended meaning

EXPR, APPLY, INVERSE, POWER, FACTORIAL, LOWLIMIT,

Both types of tags can be mixed.

MathML: an example

$$\int_0^a x^n dx$$

```
<EXPR>                                    <MROW>
  <INT/>                                    <MSUBSUP>
  <LOWLIM-                                    <MO>&int;</MO>
IT><MN>0</MN></LOWLIMIT>                       <MN>0</MN>
   <UPLIMIT><MI>a</MI></UPLIMIT>              <MI>n</MI>
   <EXPR>                                   </MSUBSUP>
     <MI>x</MI><POWER><MI>n</MI>            <MSUP>
   </EXPR>                                     <MI>x</MI>
    <BVAR><MI>x</MI></BVAR>                    <MI>n</MI>
</EXPR>                                      </MSUP>
                                            <MO>&dd;</MO>
                                            <MI>x</MI>
                                          </MROW>
```

Document Object Model (DOM)

An application programming interface (API) that allows programs and scripts to manipulate documents

- the logical structure
- the presentation

- the content

The API is:

- platform–independent
- language independent
- used with current and future versions of HTML, XML, CSS

DOM: Requirements

Manipulations:

- add, change, delete, move
- navigate

Document objects exposed to DOM:

- logical structure: (implied) elements, attributes
- presentation: style sheets, selectors, rules, properties
- content: text, images
- DTD

DOM: events

Any user's interaction raises an event:

- All elements are capable of generating events
- Interaction events, update events, and change events
- Events bubble through the structural hierarchy of the document
- Events are synchronous

Concluding remarks

A document has several structures:

- Logical structure: HTML, XML–lang
- Presentation structure: CSS, XML–style
- Semantic structure: HTML (links), XML–link
- Specific behavior: DOM
- Temporal structure: MML

DISCUSSION

Rapporteur: Dr Paul Watson

Lecture One

Dr. Quint described the mechanism by which standards are adopted by W3C. A proposed standard must be accompanied by running code. It is accepted if no W3C members object.