1998-1999 ACADEMIC TRAINING PROGRAMME

LECTURE SERIES

SPEAKER : H. LIE, World Wide Web Consortium, Oslo, N.
TITLE : Web publishing today and tomorrow
TIME : 17, 18 & 19 March, from 11.00 to 12.00 hrs
PLACE : Auditorium

ABSTRACT

The three lectures will give participants the grand tour of the Web as we know it today, as well as peeks into the past and the future. Many three-letter acronyms will be expanded, and an overview will be provided to see how the various specifications work together. Web publishing is the common theme throughout the lectures and in the second lecture, special emphasis will be given to data formats for publishing, including HTML, XML, MathML and SMIL. In the last lectures, automatic document manipulation and presentation will be discussed, including CSS, DOM and XSL.
Web publishing today and tomorrow

Håkon Lie

howcome@w3.org

Style Sheets Activity Lead

CERN, March 17-19, 1998

CERN lectures

- Wed: Grand Tour
- Thu: Content Formats
- Fri: Formatting Content

- Web Publishing
- Technical angle
- Review W3C Specifications
- Publishing strategies

The Grand Tour

- look back in history
- provide navigational aids for manoeuvring
- expand many tree-letter acronyms
- name the parts of the puzzle -- and see how they fit together
- to understand the Web, you must know the specifications
- ... and a few more things

Ladder of abstraction
The three columns

- HTTP: HyperText Transfer Protocol
- URL: Universal Resource Locators
- HTML: HyperText Markup Language

- invented by Tim Berners-Lee at CERN 1990

Column 1: HTTP

- similar to earlier protocols:
  - NNTP: Network News Transfer Protocol
  - SMTP: Simple Mail Transfer Protocol
- built on top of TCP/IP
- simple GET, PUT, POST, HEAD methods
- used for HTML, but also GIF etc.
Column 2: URL

- URL: Universal Resource Locator
- URL: Uniform Resource Locator

```
protocol | machine             | internal identifier
--------|---------------------|---------------------
```

http://www.w3.org/people/howcome

- protocol can be ftp, gopher etc.
- machine name relies on DNS

URL

- most unique concept
- popularized in techno-slang:

http://www.w3.org/people/howcome
www.w3.org/people/howcome

- heritage from UNIX, not DOS
- problem: internationalization (I18N)

http://www.w3.org/people/hakon

URI/URN

- Uniform Resource Identifier
- Uniform Resource Name

A URI can be further classified as a locator, a name, or both. The term "Uniform Resource Locator" (URL) refers to the subset of URI that identify resources via a representation of their primary access mechanism (e.g., their network "location"), rather than identifying the resource by name or by some other attribute(s) of that resource. The term "Uniform Resource Name" (URN) refers to the subset of URI that are required to remain globally unique and persistent even when the resource ceases to exist or becomes unavailable.

- URN problem not yet solved
- future-proofing terms
- URL still valid term

Column 3: HTML

```html
<H1>Headline</H1>
```
Sample HTML

<H1>Headlines are <EM>important</EM></H1>

HTML

- HTTP: hardcore network technology
- URL: design is done, or too hard
- HTML
  - everyone can invent a tag!
  - so many, that the IETF WG stalls
  - will last for a long time
  - captures human knowledge
  - <i>kulturbarer</i>

HTML timeline

- 1965: "hypertext" coined by Ted Nelson
- 1980: Tim Berners-Lee writes "Enquire" for Norsk Data @ CERN
- 1990: Robert Cailliau and Tim Berners-Lee proposes the Web, develops browser/editor @ CERN
- 1993: Marc Andreessen develops Mosaic @ NCSA, extends HTML
- 1994: Netscape is set up, adds more extensions

Shinagawa

Ladder of abstraction

semantics

HTML

presentation
Semantics vs. presentation in HTML

- classic conflict in HTML
- SGML heritage: encode structure, not presentation
- HTML 2:
  - mostly structural elements: H1, H2, P, OL, EM
  - some presentational: B, I, PRE
- HTML 3.2
  - added tables
  - intended to be structural, but used for presentation
  - FONT added as a pragmatic move
- HTML 4.0
  - presentational elements deprecated
  - style sheets pointed to as replacement
  - STYLE element, STYLE attribute added

Problems implementing HTML

- elements have been added ad hoc
- as a compromise, some of them are in specifications
- error recovery not defined
- tables need pixel-perfect rendering
- interface to Javascript not defined
- implementing the specification isn't enough

GIF

- first implemented in Mosaic
- text + images = wow!
- lossless image format
- much used banners, line drawings, digrams, screen dumps, text
- 8 bit
- plagued by patent
- animated: abcnews miningco dilbert

Ladder of abstraction
JPEG

- lossy compression
- developed for photographs
- Discrete Cosine Transform
- Examples: Shinagawa

Ladder of abstraction

semantics

JavaScript

- scripting language developed by Netscape
- initially called JavaScript
- more popular than Java
- not a close relationship
- standardized in ECMAScript
- interfaces to HTML not standardized
- reusability: low
- portability: medium

Web anno 1997

- HTML
- GIF
- JPEG
- JavaScript

World Wide Web Consortium

- W3C
- industry consortium
- 300+ members
- MIT (USA), INRIA (France), Keio (Japan)
- 50 staff
- From CERN: Tim Berners-Lee, Henrik Frystyk Nielsen, Hakon Lie, Janne Saarela
- HTML, HTTP, URL, CSS, XML, SVG, MathML, SMIL, DOM, PNG, XSL, RDF
- Not: Java, JavaScript, GIF, JPEG
Web Publishing today and tomorrow: Content Formats

Håkon Lie

howcome@w3.org

Style Sheets Activity Lead

CERN Day 2: Content Formats

Web anno 1997

- HTML
- GIF
- JPEG
- JavaScript

W3C specifications

1. W3C Working Drafts
2. W3C Proposed Recommendations
3. W3C Recommendations

PNG

- GIF + 10 years - patents + transparency = PNG
- better compression
- extension mechanism: chunks
- no animations
- supported by all current browsers

Ladder of abstraction

semantics

HTML

GIF, JPEG, PNG

presentation
CSS

- Cascading Style Sheets
- HTML describes content and structure
- CSS describes presentation

CSS sample

```html
<H1>Headline</H1>

H1 {
  font: 18pt Helvetica;
  color: red;
  margin-left: 2em;
}
```

Benefits of style sheets

- both users and authors can influence presentation
- pages are lighter, images are fewer
- text is more accessible
- esthetics

Ladder of abstraction

```
semantics

HTML
CSS
GIF, JPEG, PNG

presentation
```

DOM

- Document Object Model
- API for interacting with documents
- for use by scripts etc.
  - content
  - structure
  - style

XML

- Extensible Markup Language
- toolkit to define markup languages
• XML is only syntax, no semantics
• simplified SGML
• make your own tags!

XML Sample

<article>
  <headline>CERN finds small things</headline>
  <introduction><location>Geneva</location>At the European Insitute...</introduction>
  <paragraph>But, as scientists quickly point out, finding small things is hard.</paragraph>
</article>

• end tags can't be omitted
• tag names are case-sensitive
• no semantics nor presentation
• style sheets always needed

XML pros and cons

• pro:
  • you can make your own tags
  • draconian
  • many tools support XML
• con:
  • you can make your own tags
  • draconian
  • few browsers support XML

HTML's CLASS attribute

<BODY CLASS="article">
  <H1 CLASS="headline">CERN finds small things</H1>
  <P CLASS="introduction">
    <SPAN CLASS="location">Geneva</SPAN>At the European Insitute...
  </P>
  <P CLASS="paragraph">But, as scientists quickly point out, finding small things is hard.</P>
</BODY>

• HTML can contain all semantics
• compatible with current browsers

When to use XML

• server-side applications
• non-document data
• when you can establish global semantics
XHTML

- HTML4 recast in XML
- end tags can't be omitted
- tag names will be lower-case?

Content-Type: text/xhtml

MathML

- finally!
- written in XML
- has both semantics and presentation:
  - "content" elements (describes meaning)
  - "presentation" elements (describes notation)

"A fundamental challenge in defining a mathematics markup language for the Web is reconciling the need to encode both the presentation of a mathematical notation and the content of the mathematical idea or object which it represents."

"Twenty-eight of the MathML tags describe abstract notational structures, while another seventy-five provide a way of unambiguously specifying the intended meaning of an expression."

MathML presentation elements

```
a-b
```

```
<apply>
  <minus/>
  <ci>a</ci>
  <ci>b</ci>
</apply>
```

MathML content elements

```
a-b
```

```
<apply>
  <minus/>
  <ci>a</ci>
  <ci>b</ci>
</apply>
```
Ladder of abstraction

semantics

MathML
HTML
CSS, MathML
GIF, JPEG, PNG

presentation

SVG

- Scalable Vector Graphics
- vector graphics for the Web
- beyond HTML's <BR> element
- written in XML

<circle style="fill: blue; stroke: red"
cx="200" cy="200" r="100"/>

- rect
- circle
- ellipse
- polyline
- polygon
- pie slice
- text

Ladder of abstraction

semantics

MathML
HTML
CSS, MathML, SVG
GIF, JPEG, PNG

presentation

Resource Description Framework (RDF)

- metadata
- data about data (e.g. library catalog)
- the Web is machine-readable, but not machine-understandable
• will help automation
  • better search engines
  • automatic exchange of privacy information
  • content rating
  • intellectual property rights
• Web of trust
• "BTW, file my taxes!"

RDF example

"Håkon Lie is the author of http://www.w3.org/people/howcome"

The sentence has the following parts

• Resource (Subject): http://www.w3.org/people/howcome
• Property (Predicate): Author
• Value (Object): Håkon Lie

Resource + Property + Value of property = Statement

RDF coded in XML

```xml
<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/TR/1999/PR-rdf-syntax-19990105#"
  xmlns:s="http://description.org/schema/"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.w3.org/2000/10/rdf-schema.xsd"
>
  <rdf:Description about="http://www.w3.org/people/howcome"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
    <s:Creator>Håkon Lie</s:Creator>
  </rdf:Description>
</rdf:RDF>
```

• XML is only one of several possible syntaxes

http://www.w3.org/people/howcome { creator: Håkon Lie }

Ladder of abstraction

```
semantics

<table>
<thead>
<tr>
<th>RDF</th>
<th>MathML</th>
<th>HTML</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSS</td>
<td>MathML</td>
<td>SVG</td>
</tr>
<tr>
<td>GIF</td>
<td>JPEG</td>
<td>PNG</td>
</tr>
</tbody>
</table>
```

Web Publishing today and tomorrow: Formatting Content
Håkon Lie

howcome@w3.org

Style Sheets Activity Lead

CERN Day 3: Formatting Content

Formatting content? Why?

- humans need formatting
- centuries of calligraphy and typography
- 1993: the Web looked boring
  - gray backgrounds
  - no runaround text
  - no fonts
- 1997: the Web is a disaster zone
  - pages made for 15" monitors
  - text inside images don't print well
  - typographical expressiveness low

Why style sheets?

- provides typographical expressiveness
- content can remain device-independent
- accessibility
- quicker downloads

Case study: Microscape

- Microscape = Microsoft + Netscape home pages
- 40 GIF images
- 22 images replace by HTML+CSS
- CSS more important than PNG and HTTP 1.1

CSS sample

```html
<H1>Headline</H1>

H1 {
    font: 18pt Helvetica;
    color: red;
    margin-left: 2em;
}
```

CSS

demo
- Cascading Style Sheets
• 1996: level 1
  • fonts
  • color
  • white space
• 1998: level 2
  • aural renderings
  • downloadable fonts
  • positioned elements
  • bi-directional text
  • tables

**CSS Implementations**

• 96: Microsoft IE 3
• 97: Netscape N 4, MS IE 4
• 98:
  • Netscape 5 (codename Gecko)
  • Microsoft IE 5
  • Opera 3.5

**How to get started**

• first CSS implementations poor
• hard to write style sheets that work on all browsers
• point to W3C Core Styles
• http://www.w3.org/StyleSheets/Core

**Improving CSS implementations**

• post-recommendation activities
• CSS1 Test suite
• W3C Core Style Sheets
• CSS Validator
• http://www.w3.org/Style
• http://www.webstandards.org

**CSS2 Media-specific style sheets**

```css
BODY {
  color: black;
  background: white;
}
```

```css
@media tv {
  BODY {
    color: white;
    background: black;
  }
}
```

**Floating elements vs tables**
- HTML tables commonly used for layout
- Floating element is more scalable
- demo

**XSL**

- Extensible Stylesheet language
  1. tree transformation language
     - transform from XML to XML
     - unlike DSSSL: declarative
     - will be called XTL
  2. set of formatting objects
  3. will be called XFO
- XTL Recommendation June 99

**XTL example**

```xml
<xsl:template match="Heading1">
  <H1>
    <xsl:apply-templates/>
  </H1>
</xsl:template>

input: <Heading1>Overskrift</Heading1>
output: <H1>Overskrift</H1>
```

**XTL**

- server side transformations XML -> HTML
- automatically generate table of contents
- XTL is declarative
- through DOM, scripts can do the same

**XFO?**

- all CSS1/CSS2 properties will be usable
- focus on print
- new functionality = CSS3?

**Web publishing conclusion**

- capture semantics
- preserve semantics as long as possible
- publish media-independent content
- use media-specific style sheets
- be ready for new Web devices
- have fun!