



XHTML: An Introduction

Prof. Ing. Andrea Omicini

Ingegneria Due, Università di Bologna a Cesena

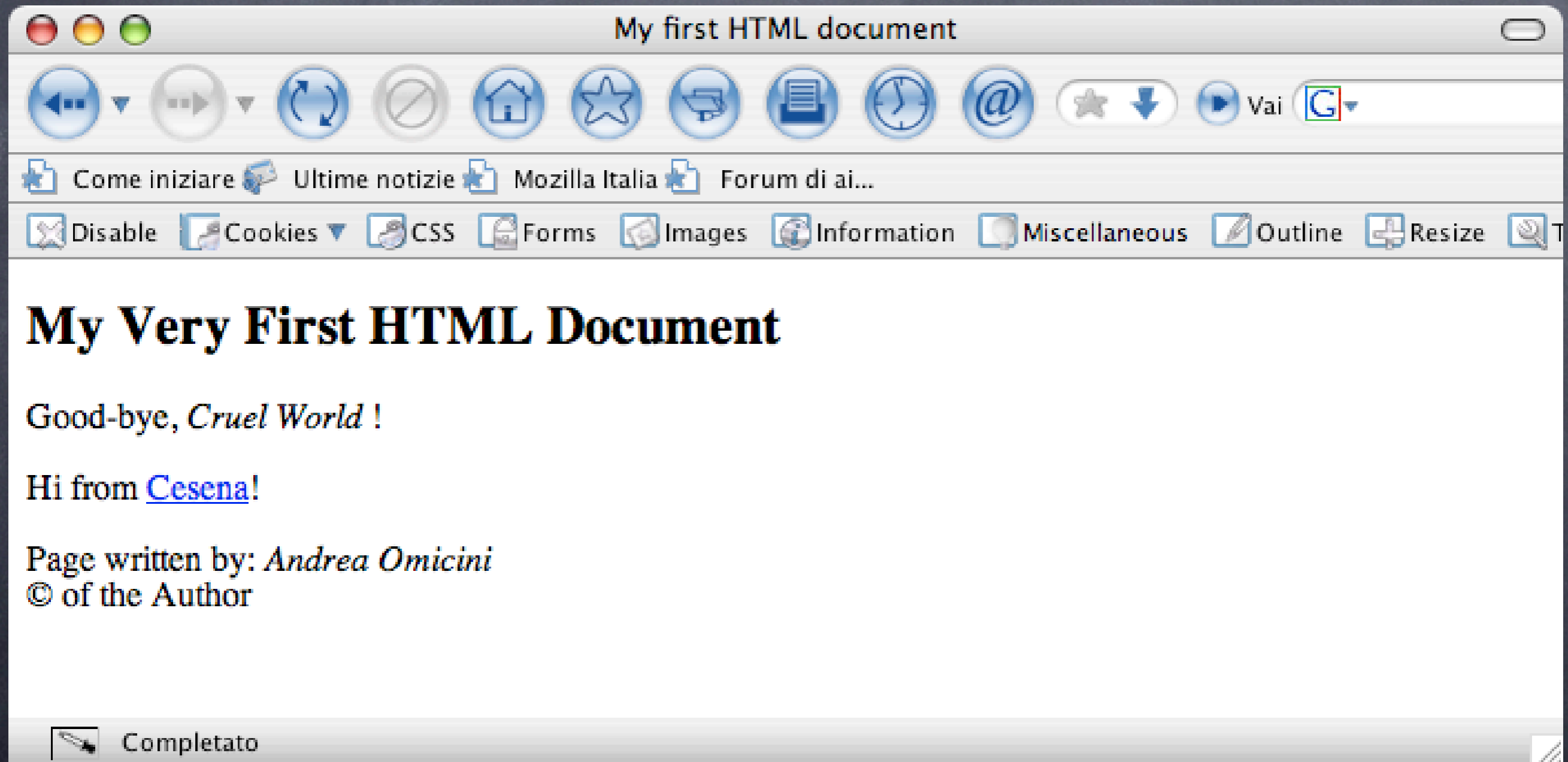
andrea.omicini@unibo.it

2005–2006

Good-bye Cruel World!

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
<head>
    <title>My first HTML document</title>
</head>
<body>
<h2>My Very First HTML Document</h2>
<p>
    Good-bye, <i>Cruel World </i>!
</p>
<!-- "Hello World" did not seem enough -->
<p>
    Hi from <a href="http://www.ing2.unibo.it/">Cesena</a>!
</p>
<p>
    Page written by: <cite>Andrea Omicini</cite>
    <br />
    &#169; of the Author
</p>
</body>
</html>
```

We Obtain...



What is HTML?

- It is a markup language
 - allows you to annotate text, and to embody annotations in along with text in a document
 - annotations label text properties
 - e.g., printing properties as annotations, in order to separate them from content
 - SGML subset
 - Standard Generalized Markup Language
- A family of standards
 - W3C: consortium in charge of Web standards
 - <http://w3c.org/Markup>
 - Develops over time
 - either official or proprietary extensions
 - proposals, drafts, releases and recommendations

Versions

- From 1.0, 2.0, 3.2, 4.0 to 4.01
 - HTML 4.01 is the last recommendation
 - <http://www.w3.org/TR/html4/>
- XHTML 1.0 current standard
 - defined based on HTML 4.01
 - as more or less its XML-compliant version
 - <http://www.w3.org/TR/xhtml1/>
- XHTML 2.0 still ongoing work
 - <http://www.w3.org/TR/xhtml2/>
 - 27/5/2005: 7th public Working Draft of XHTML 2.0 has been published
 - <http://www.w3.org/TR/2005/WD-xhtml2-20050527/>



XHTML™ 2.0

W3C Working Draft 27 May 2005

This version:

<http://www.w3.org/TR/2005/WD-xhtml2-20050527>

Latest version:

<http://www.w3.org/TR/xhtml2>

Previous version:

<http://www.w3.org/TR/2004/WD-xhtml2-20040722>

Diff-marked version:

<xhtml2-diff.html>

Editors:

Jonny Axelsson, Opera Software

Mark Birbeck, x-port.net

Micah Dubinko, Invited Expert

Beth Epperson, Websense

[Masayasu Ishikawa](#), W3C

[Shane McCarron](#), [Applied Testing and Technology](#)

Ann Navarro, WebGeek, Inc.

[Steven Pemberton](#), CWI (HTML Working Group Chair)

This document is also available in these non-normative formats: [Single XHTML file](#), [PostScript version](#), [PDF version](#), [ZIP archive](#), and [Gzip'd TAR archive](#).

From SGML to HTML

- SGML is a very intricately markup language
 - Web needs just a subset of it
- SGML is also a meta-markup language
 - can be used to define other markup languages
 - by defining their own DTD (Document Type Definition)
- HTML is defined via a SGML DTD
 - <http://www.w3.org/TR/html4/sgml/dtd.html>

HTML / XHTML Document

- Suffix .html (or .htm, for windoze fault)
 - Text file
- How can I create my own HTML file?
 - with any text editor, saving text with .html suffix
 - with any word processors allowing "Save as Text"
 - with any web-page creation tool
 - Composer, Dreamweaver, etc.
 - please no FrontPage
- How does computer see the HTML file?
 - in the same way as we do?
 - different levels of "perception"
 - OS, editors, browsers, ...

Elements & Tags

- An HTML document contains
 - **elements** & sections delimited by **tags**
- Generally, tags delimit start & end of a section / element
 - `<tag>section or element</tag>`
 - It is obviously of help to learn the main HTML tags
- Tags may contain further specifications called **attributes**
 - some of them **required**, some **optional**
 - ``
 - src is mandatory
 - no closing tag, `/>` is used
- `<p>` is a tag, `<p>Paragraph</p>` is an element

Some Details

- White spaces have no meaning
 - if not within strings
- HTML is not case sensitive
 - `<p>` or `<P>`, it is the same
- this does not hold for XHTML!
 - `<p>` is correct, `<P>` is wrong

Types of Tags

- 👁 Section tag
- 👁 Header tag
- 👁 Content tag
- 👁 Styling tag
- 👁 Empty elements
- 👁 Anchor / Hyperlink tags

Section Tags

- 👁️ dividing HTML document in sections
- 👁️ root tag
 - 👁️ `<html>` starts HTML document
 - 👁️ may not be the beginning of the HTML file...
 - 👁️ `</html>` ends it
 - 👁️ while the file might go on...
- 👁️ 2 sections: Header & Body

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

Header tags

- within the header, between `<head>` and `</head>`
 - not displayed directly by the browser
- main header tags
 - `<title>` defines page title
 - in the title bar of the browser window
 - `<meta />` carries meta-information on the document content
 - e.g.: `<meta name="author" content="Andrea Omicini" />`
 - like a comment, but can be referenced and used

Content tags

- ◉ within the body, between `<body>` and `</body>`
 - ◉ used by browser for display
- ◉ most of the useful tags
 - ◉ `<p>` for paragraphs
 - ◉ `<table>` for tables
 - ◉ `<h1>` for 1st-level headers
 - ◉ `<h2>`, `<h3>`, ... next levels
 - ◉ comments
 - ◉ `<!-- this is a comment -->`

Styling tags

- two kinds
 - based on the content nature
 - based on formatting
- content-based: examples
 - `<blockquote>` contains a block for a quotation
 - `<cite>` contains a reference to a citation
- format-based: examples
 - `` bold, `<i>` italic
- sometimes no differences in display by browsers
 - ``, `<cite>`, `<i>`, `<dfn>`
 - but the source shows the differences in markup
 - that could be used anyway for some reasons

Empty tags

- 👁 In XHTML
 - 👁 `
` line break
 - 👁 `<hr />` horizontal rule
 - 👁 `` inline image
- 👁 In HTML , `
` & `<hr>` are ok
 - 👁 in XHTML they should be "closed" somehow
- 👁 Pay attention to attributes!
 - 👁 required & optional
 - 👁 e.g., attribute `src` in `img` is required
 - 👁 take a close look to specifications
 - 👁 check when needed
 - 👁 exploit tools!
 - 👁 along with their embedded knowledge

Anchor / Hypertext tags

- 👁 tag `<a>` for both
 - 👁 "anchor" denotes portions of a document
 - 👁 to be directly referred to with #
 - 👁 "hypertextual reference" denotes other docs
 - 👁 or portions of them
 - 👁 obviously contains an URL
- 👁 `...`
- 👁 `...`
 - 👁 relative / absolute URL
 - 👁 `` denotes an anchor within an href
- 👁 Pay attention to quotes!

Limits of HTML (1)

- Content intermixed with presentation
 - from 1.0 to 4.01 things have improved
 - but too many biases from browsers
 - to be absolutely AVOIDED
 - in general
 - here in this course for sure :)
- Not "well-formed"
 - as XML is instead
 - too much forgiving
 - elements can be interleaved, tag can be wrong, closing tags or attributes may be missing, etc...

Limits of HTML (2)

- It is more a sort of "structural" markup language
 - describes text structure
 - structural markup
 - more than the nature of content
 - descriptive / semantic markup
 - not easy to adapt to the different natures of media
- That is why we are moving to XML
 - through XHTML
- Goal: a language
 - disciplined and easy to check
 - powerful but simple
 - descriptive

XML in short

- Extensible markup language
 - to define markup languages
- XML application
 - a markup language defined via XML
 - XHTML is an XML application
- Essential remark: XML has no predefined tag / elements
 - anyone can define tags and structures that better fits the chosen contents

Fundamental Parts of a XML Document

1. XML Document (properly said)
 - 👁 content built according to XML rules
2. Document Type Definition (DTD)
 - 👁 which tags and their meaning
3. Style Sheet
 - 👁 for presentation

Benefits of XML

- Portable
 - text format, any platform and many applications t
oread & write it
- Configurable / Extensible
 - anybody can define his/her own markup language
- Self-descriptive
 - an XML document is self-contained: presentation,
meaning, data & their structure

XHTML = HTML + XML

- XHTML defined using XML as a meta-language
 - HTML defined instead in SGML
- vocabulary taken from HTML, syntax from XML
 - backward compatibility
 - in particular, in the "human legacy"
 - XML properties
 - well-formed, not error-prone, extensible via XML

XHTML - HTML = ?

👁 XML Prologue

- 👁 first element of the XHTML document

- 👁 `<?xml version="1.0" encoding="UTF-8"?>`

- 👁 like corresponding `<meta />` for old browser

👁 Document type declaration

- 👁 what is the document DTD?

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

- 👁 address, or embedded DTD

- 👁 both before `<html>` tag

Other differences

- 👁 Recommendation: define the namespace

- 👁 to give meaning to tags

- 👁 we could write

- `<http://www.w3.org/1999/xhtml:p>` for `<p>` tag

- 👁 and then the same all the others...

- 👁 however, it seems easier to write

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

- 👁 also, we could add our namespaces

- 👁 and declare the languages

- 👁 Other

- 👁 case sensitivity

- 👁 full nesting

- 👁 required elements: `<head>`, `<body>`, `<html>`, `<title>`

What should we learn from the lab activity?

- 👁 Structure of the HTML / XHTML document
 - 👁 header, body, and their content
- 👁 Inline elements, comments, lists, special chars
- 👁 Attributes
 - 👁 shared by most elements
 - 👁 attributes to affect presentation
- 👁 Anchors / Hypertextual references
- 👁 Images
- 👁 Tables
- 👁 Forms
- 👁 Frames, perhaps