Internet content

HTML
SGML
CSS
XML
XHTML
MIME
HTTP

Objectives

- What HTML is, what its origins are, and where one can find information about it
- Next generation HTML: CSS, XHTML
- Describing internet data: XML
- Understanding how different types of content are dealt with in the Internet (MIME)
- HTTP, internet protocols, proxies

HTML & SGML

- HyperText Markup Language
- Markup:
  - Classic usage: editing marked up text
  - Correcting, data description and other usages
- Markup languages have a long history
  - troff, nroff (for unix man pages), runoff
  - TeX, LaTeX, excellent to write books with (Knuth) and slides (eg. these slides)
  - SGML, the origin of HTML, looks like today's XML. SGML and XML describes data

SGML/XML-example

```
<email>
   <sender>
      <person>
         <id>serafim@csc.kth.se</id>
         <christianname>Serafim</christianname>
         <familyname>Dahl</familyname>
      </person>
   </sender>
   <reciever>
      <person>
         <distributionlist>DD1335;gruint10@kth.se</distributionlist>
      </person>
   </reciever>
   <contents>
      It's ugly, isn't it?
   </contents>
</email>
```

SGML/XML-example . . .

- The example describes just data
- It says nothing about presentation (color, font, alignment, . . .)
- All text is inside a `<tag>text</tag>`
- There is a separate document that specifies what tags are allowed in the SGML document, in what order are they allowed, etc. Can be used to validate the SGML

DTD, Document Type Definition

```
<!doctype email[
<!element email (sender,reciever,contents)>
<!element sender (person)>
<!element reciever (person)+>
<!element person (distributionlist) |
                 (christianname, middlename?,familyname)>
<!element (christianname, middlename?,familyname)
          (#PCDATA)>
<!element distributionlist (#PCDATA)>
<!element contents (#PCDATA)>
]>
```


HTML

- Describes how webpages are visualised
- A “web browser” reads the description and interprets it or (when there is no description) uses a default interpretation
- early version 1965 in Douglas Englebart’s “oNLine System”, NLS
- Tim Bernes-Lee at CERN made ENQUIRE in 1980, which developed into HTML and present presentation techniques.
- HTML is developped and maintained by W3C (World Wide Web Consortium)
  - HTML, CSS, XHTML, XML
  - Most common is HTML 4.0 (default)
  - Latest version is XHTML 1.1

HTML . . .

```
<html><head><title>HTML</title></head><body>
<!- the line above may be omitted ->
   <h1>HTML</h1>
   <p>This is a short presentation of <b><u>HTML</u></b>. Its
      main points:</p>
   <ul>
      <li>Unlike SGML and XML, HTML describes how data is
          <i>presented</i>, not what the data <i>is</i>. It is thus
          an editing markup, much like
          <a href="http://www.tex.ac.uk/cgi-bin/texfaq2html">TeX</a>
          <ul>
             <li>There can be text outside any tag, and though it
                 won't be validated, it will &quot;work&quot;</li>
             <li>You can't write a validator document</li>
             <li>If a HTML document is invalid e.g. by not having
                 correct tag order, or missing tags, it will be presented
                 anyway. A closing tag is missing right here
          </ul></li>
      <li>As in other markup languages, some tags can only appear
          inside other tags (e.g. &lt;li&gt; can only appear inside
          &lt;body&gt;)</li>
          <li> A text fragment in a document can link to other
               documents, or to a specific place in the document.</li>
   </ul>
</body></html>
```

will look like:

HTML

This is a short presentation of **HTML**. Its main points:

- Unlike SGML and XML, HTML describes how data is *presented* not what the data *is*. It is thus an editing markup, much like TeX
  - There can be text outside any tag, though this won't be validated, it will "work"
  - You can't write a validator document
  - If a HTML document is invalid e.g. by not having correct tag order, or missing tags, it will be presented anyway. A closing tag is missing right here
- As in other markup languages, some tags can only appear inside other tags (e.g. can only appear inside <body>)
- A text fragment in a document can link to other documents, or to a specific place in the document.

XHTML

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!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">
<head><meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1" />
  <title>XHTML</title>
</head>
<body><h1>XHTML</h1>
  <p>This is a short presentation of <b><u>XHTML</u></b>.  Its main points:</p>
  <ul>
    <li>Unlike SGML and XML, XHTML describes how data is
      <i>presented</i> not what the data <i>is</i>. It is thus an editing
         markup, much like  <a href="http://www.ctan.org/">TeX</a>
      <ul><li>There can be text outside any tag, though this won't be validated,
            it will &quot;work&quot;</li>
        <li>You can't write a validator document</li>
        <li>If an XHTML document is invalid there will be an error message
            A closing tag is missing right here
      </ul></li>
    <li>As in other markup languages, some tags can only appear inside
       other tags (e.g. <li> can only appear inside <body>)</li>
    <li>A text fragment in a document can link to other documents, or to a
       specific place in the document.</li>
  </ul>
</body>
</html>
```


XHTML - result

```
XML Interpreter error: mismatched tag. Expected: </li>.
Adress: http://www.csc.kth.se/~serafim/02-internet-content.xhtml
Radnummer 22, Kolumn 9:
      </ul>
------^
```


HTML tags

- HTML reference: `http://www.htmlhelp.com/reference/html40/`
- HTML referece: `http://www.webreference.com/authoring/languages/html/`
- XHTML referece: `http://www.webreference.com/authoring/languages/xhtml/`
- Also on `http://www.w3schools.com/` where there are links to other places
- Organizational list of HTML tags `http://www.htmlhelp.com/reference/html40/olist.html`

HTML tags . . .

- Parts of an HTML document:
  - the header contains general information about the document e.g. the title, author, document generator, . . .
  - the body contains the document content or
  - a frameset that describes a set of frames
  - If the header is missing, the whole content is interpreted as body
- The tag `<a>` with the attribute `href` denotes a link. The link may be absolute or relative. It is good practice to use relative links to pages in the same site, because then moving the site to another server or directory is easy
  - `<a href="document-in-same-directory.html">...`
  - `<a href="directory/dokument.html">...`
  - `<a href="../another-directory/another-doc.html">...`
  - `BASE` can be used to indicate the base for all relative links

HTML tags . . .

- Other useful head tags: `title` (most used)
- Later: `meta` (used to simulate HTTP headers)
- `style` used to change the outlook of a document eg. by using *cascading style sheets* (`CSS`)
- In the body: `<p>...</p>` denotes a paragraph. If there's no tag around a text fragment, the default is `<p>...</p>`. Use `<br>` or `<br />` for a line break.
- Anchors: `<a name="here">` marks a place in the document that can be accessed using `documentname#here` like `http://www.csc.kth.se/utbildning/kth/kurser/DD1335/index.php#start`
- `<h1> </h1> ... <h6> </h6>` are heading levels
- If you want spaces and newlines to matter, use: `<pre> </pre>`
- Text appearance can be changed like in text editors with `<b>`, `<i>`, `<font>`, `<big>`, `<small>`, `<sub>`, `<tt>`, `<s>`. The modern choise is to use `style`, mandatory in XHTML, wise even in HTML.
- Tables `<table>`, table rows `<tr>`, table headers `<th>`, table cell `<td>`

HTML tools

On the web there are many application programs that can be used to check your (X)HTML documents, even to upgrade them to later standards

- `http://www.htmlhelp.com/tools/validator/`
- `http://sourceforge.net/projects/tidy/`
- `http://jigsaw.w3.org/css-validator/`
- `http://www.thefreecountry.com/webmaster/htmleditors.shtml`

There are also development systems for HTML/CSS/PHP/EJB offering extensive help and support when programming for the web, on Linux e.g.

- BlueFish
- Quanta+
- Eclipse (for all platforms and most languages)
- Netbeans (for all platforms and many languages)

Style / Cascading Style Sheets (CSS)

You can redefine the style of various HTML elements. The style can be defined in the HTML file:

```
<html>
  <head>
    <style type="text/css">
      h2 {text-decoration:overline} h4 {text-decoration: line-through}
      p {text-decoration: underline} a {text-decoration: none}
    </style>
  </head>
  <body>
    <h2>A level 2 heading</h2>
    <h4>A level 4 heading</h4>
    <p>A paragraph</p>
    <p>Another paragraph <a href="http://w3schools.com">
       with a link</a></p>
  </body>
<html>
```

CSS . . .

A level 2 heading

~~**A level 4 heading**~~

A paragraph

Another paragraph with a link

CSS . . .

Style can also be defined

- in a separate CSS file indicated in a ”link” attribute in the head section of the HTML file:

```
<link rel="stylesheet" type="text/css" href="mystyle.css">
```

- or directly in the HTML code:

```
<p style="color: sienna; margin-left: 20px">
   This is a paragraph</p>
```


XML

- XML is a way of describing data, according to a DTD (like SGML)
- There are no predefined XML tags (like in HTML). One has to describe ones own using e.g. a DTD
- XML doesn’t do anything in itself but data from one XML document can be presented in an indefinite number of ways
- XML can be used to exchange data, to express the configuration of software in a rich, hierarchical manner or even as program source code
- Example of an XML application: RSS (really simple syndicating). See eg.
  `http://www.mozilla.org/products/firefox/live-bookmarks.html`

XML: example

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<note>
   <date>2010-01-19</date>
   <time>08:23 GMT+1</time>
   <to>Serafim</to>
   <from>Carina</from>
   <heading>Reminder</heading>
   <body>Don’t forget our lunch meeting today!</body>
</note>
```

XML: example . . .

This XML-file does not appear to have any style information associated with it. The document tree is shown below.

```
-<note>
   <date>2009-01-19</date>
   <time>08:23 GMT+1</time>
   <to>Serafim</to>
   <from>Carina</from>
   <heading>Reminder</heading>
   <body>Don't forget our lunch meeting today!</body>
 </note>
```


XHTML

- eXtensible HTML
- XHTML is designed to replace HTML
- XHTML 1.0 is almost identical to HTML 4.1
- A stricter and cleaner version
  - All tags must be closed `<br>` becomes `<br/>`. Better with `<br />` as old browsers understand them.
  - All documents must be well-formed (text is illegal)
  - There should be no text outside tags
- XHTML is HTML defined as an XML application: `http://www.w3c.org/TR/xhtml1/dtds.html`

MIME: Multimedia Internet Mail Extension

- An open standard describing how multimedia is sent via email (initially) HTTP request, response, etc
- Describes how the parts of the content are organized. A part can contain other parts, . . .
  Globally/parts/subparts/. . .
- Describes a type for the data sent, for example
  - text
    - plain, html
  - news
  - postscript, pdf, doc
  - zip
  - image
    - jpeg, tiff, gif, . . .
  - audio
  - video
    - mpeg, quicktime, wmv . . .

HTTP: HyperText Transport Protocol

- Standard that describes how a web client (mostly a browser) and server communicate to exchange data
- Uses MIME to encode data both from the browser to the server (request) and back (response)
- Uses TCP/IP for data transfer
- To get the `/utbildning/kth/kurser/DD1335/gruint09/index.html` page the client sends a request, like `GET /utbildning/kth/kurser/DD1335/gruint09/index.html HTTP/1.1`
- See http://www.w3c.org/Protocols

HTTP, request

- Command (`GET` or `POST`)
  `GET /utbildning/kth/kurser/DD1335/index.php HTTP/1.1`
- Headers (name:value pairs)
  `Host: www.csc.kth.se`
  `Accept: */*` accept any `MIME` type
- empty line
- Content (in the declared content-type, nothing in the case of GET)
- If it is a `POST` request, content is sent too

HTTP, response

- Command (mostly `OK`)
  `HTTP/1.1 200 OK`
- Headers (for example content-type, MIME)
  `Date: Fri, 12 Jan 2009 00:11:31 GMT`
  `Server: Apache/2.2.3 (Fedora) PHP/5.1.6`
  `mod_perl/2.0.1 mod_ssl/2.0.54 OpenSSL/0.9.8b`
  `Transfer-Encoding: chunked`
  `Content-type: text/html`
- empty line
- content (in in our case the file
  `/public/www-csc/utbildning/kth/kurser/DD1335/index.php`
  at CSC)

HTTP, proxies

- Some organizations wish to limit the HTTP accesses from their network to other networks
- To do that, all browsers will have to connect to a proxy running on a host inside the local network. Other connections are restricted by a firewall
- The proxy 'looks' like a normal HTTP server, but it doesn't actually host any page. It is not limited by the firewall
- The proxy will connect to the desired host on the Internet
- The proxy can decide to block connetion to some 'banned' host
- The proxy can also do caching: if a page is requested by a lot of people in the local network, it will be served faster
- The proxy adds special headers to the HTTP response
- Some networks only require proxy for port 80. Links to http://host:8080 may still work
- Proxies can be defined for many other protocols