COMP519 Web Programming
Lecture 2: HTML (HTTP and HTML5 Basics)
Handouts

Ullrich Hustadt

Department of Computer Science
School of Electrical Engineering, Electronics, and Computer Science
University of Liverpool

Contents

Web

World Wide Web [New]

An infrastructure that allows to easily develop, deploy, and use distributed systems

Distributed systems

A system in which components located on networked computers communicate and coordinate their actions by passing messages in order to achieve a common goal

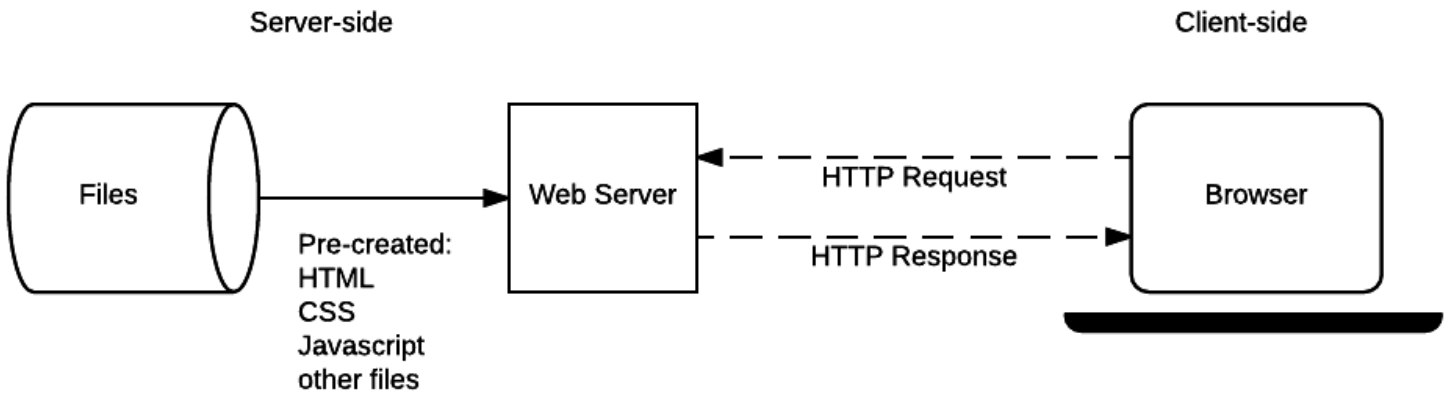
The web uses the Hypertext Transfer Protocol to communicate

(Communication) protocol

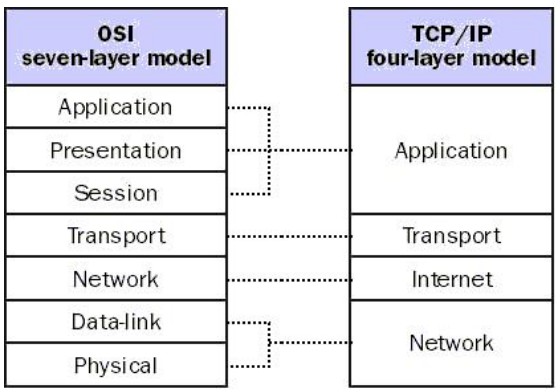
A defined system that allows two or more entities to transmit information via any kind of variation of a physical quantity
It defines the rules, syntax, semantics and synchronization of communication

HTTP

- Web clients (web browsers) and web servers use HTTP (Hypertext Transfer Protocol) to communicate with each other



- More generally, HTTP is an application-layer protocol for distributed systems



HTTP: History

- 1991 – HTTP 0.9
  first documented version of the protocol
- 1996 – HTTP/1.0
  first version of HTTP that was an Internet Engineering Task Force (IETF) informational RFC (RFC 1945)
  HTTP 0.9 and HTTP/1.0 require a separate TCP/IP connection for every resource request
- 1997 – HTTP/1.1
  first version of HTTP that was an Internet Engineering Task Force (IETF) formal standard (RFC 2068)
  HTTP/1.1 can reuse a TCP/IP connection to request several resources from the same server
- 1997-2014
- 2015

HTTP: History

- 1991 – HTTP 0.9
- 1996 – HTTP/1.0
- 1997 – HTTP/1.1
- 1997-2014
  Minor improvements and clarifications of HTTP/1.1 are developed
- 2015 – HTTP/2
  - Major revision of HTTP with focus on efficiency and privacy improvements
  - HTTP/2 allows a server to push resources to client even before they are requested
  - HTTP/2 puts more emphasis on encrypted connections

HTTP Requests

Browser/Client ⟶ Server

```
GET /index.html HTTP/1.1
Host:  www.example.com
```

Browser/Client ⟵ Server

```
HTTP/1.1 200 OK                                          Header
Date:  Mon, 24 Sep 2018 22:38:34 GMT
Content-Type:  text/html; charset=UTF-8
Content-Encoding:  UTF-8
Content-Length:  138
Last-Modified:  Wed, 10 Jan 2018 23:11:55 GMT
Server:  Apache/2.4.34 (Unix) (Red-Hat/Linux)
ETag:  "3f80f-1b6-3e1cb03b"
Accept-Ranges:  bytes
Connection:  close

<html>                                                   Message body
<head>
<title>An Example Page</title>
</head>
<body>
Hello World, this is a very simple HTML document.
</body>
</html>
```

Wikipedia Contributors: Wikipedia, The Free Encyclopedia, 16 September 2018 23:26
https://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol. [accessed 13 Sep 2017]

Character Encodings

- Computers operate on bits (0/1) and sequences of bits
- To store a text, it needs to be encoded as a sequence of bits
  To retrieve a text, a sequence of bits needs to be decoded back to a sequence of characters
- Early examples of such encodings are
  - 7-bit ASCII (American Standard Code for Information Interchange)
  - 8-bit ANSI (American National Standards Institute)
  - 8-bit Windows-1252
  - 8-bit Mac OS Roman
- However these allow to encode at most 256 characters
  $\rightsquigarrow$ the languages of the world contain many more characters

Character Encodings

- UTF-8 is a modern solution to this problem:
  (Almost?) every known character is mapped to a sequence of 1x8 bits to 4x8 bits
- Within UTF-8, ANSI characters retain their encoding

Char	ASCII	ANSI	UTF-8	Mac OS Roman
a	1100001	01100001	01100001	01100001
â		11100010	11100010	10001001
ä		11100100	11100100	10001010
α			11001001:10100011	

- When two systems exchange texts, then they need to know / agree which encoding they are using
  $\rightsquigarrow$ A HTTP header uses ASCII
  $\rightsquigarrow$ A HTTP message body can use an arbitrary encoding

Hypertext and HTML

- The Hypertext Markup Language is the language for specifying the static part of a web page / elements of an interface
- Hypertext documents contain links to other hypertext documents, creating an associative trail that readers can choose to follow
- Markup is a general term for special symbols (tags) that are added to plain text to provide additional information about document structure, content type, formatting, etc
- The terms 'hypertext' and 'hypermedia' were coined by Ted Nelson in 1963 as part of a model he developed for creating and using linked content
- The idea of hypertext is attributed to Vannevar Bush who in in 1945 described a hypothetical hypertext device called Memex in a magazine article

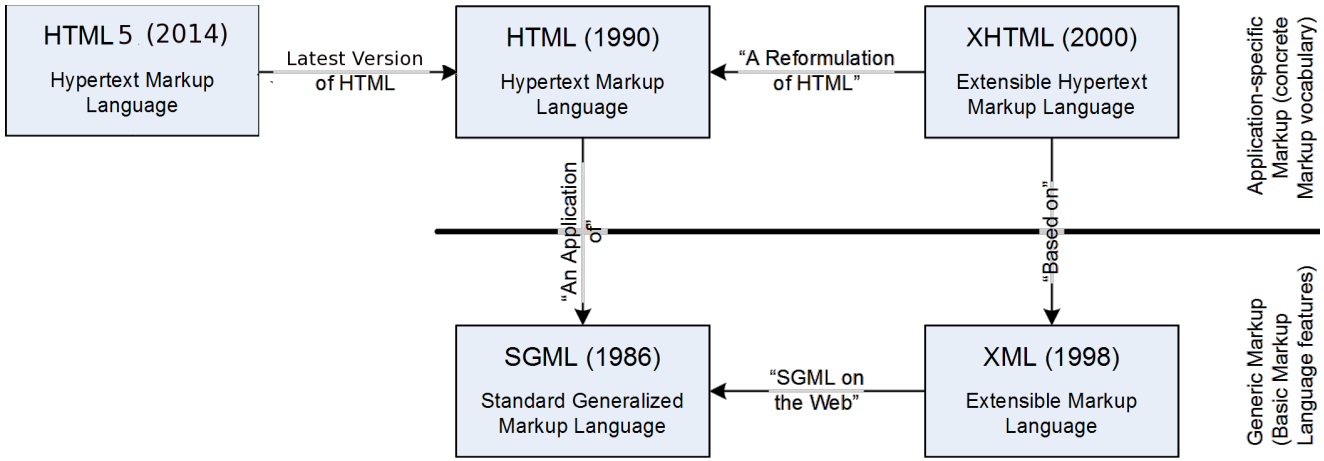
HTML: Chronology

- 1989, Berners-Lee HTML 1
  Very basic, limited integration of multimedia added in 1993,
  web browser Mosaic supported many additional features
- 1994, IETF HTML 2.0
  Tried to standardize these additional features, but during 1994–96,
  web browsers Netscape and IE supported many new, divergent features
- 1995, IETF HTML 3.0
  Proposed, but never received approval
- 1996, W3C HTML 3.2
  Again attempted to unify all features into a single standard
  but also dropped some tags that were in HTML 2.0

HTML: Chronology

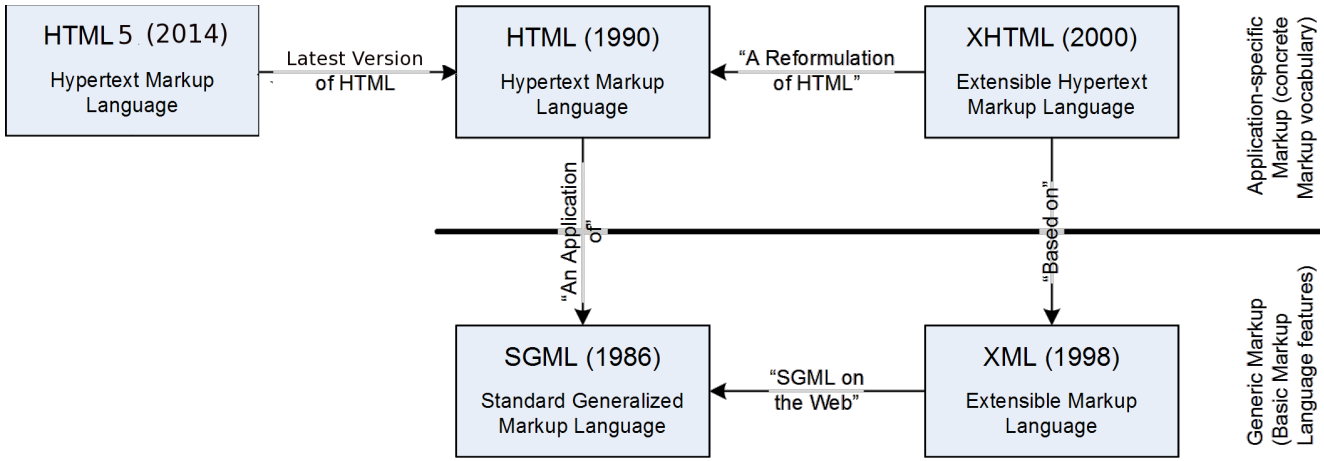
- 1997, W3C HTML 4.0
  Tried to discourage the use of 'frames', dropped Netscape visual tags, and introduced CSS; defined three variants:
  - Strict: Deprecated elements are forbidden
  - Transitional: Deprecated elements are allowed
  - Frameset: Frames are allowed
- 1999, W3C HTML 4.01
  Minor changes, the three variations are maintained
- 2000, ISO 'ISO HTML'
  ISO/IEC 15445:2000, based on HTML 4.01 Strict

HTML: Chronology



- 2000, W3C XHTML 1.0
  Reformulation of all three HTML 4.01 variations in XML
  Unlike HTML, anyone can define their own tags and attributes
  Unlike HTML, XHTML requires strict adherence to coding rules
- 2001, W3C XHTML 1.1
  Based on XHTML 1.0 Strict, introduces modules

HTML: Chronology



- 2014, W3C HTML5
  Shifts the focus from 'semantically describing scientific documents' to 'supporting web applications'
- 2016, W3C HTML 5.1
  Adds features for more responsive web apps and improved navigation
- 2017, W3C HTML 5.1 2nd Edition
- 2018, W3C HTML 5.2

Elements, Attributes and Values

- The HTML5 specification defines a set of elements, attributes, and attribute values and their meanings (semantics)
  (there are more than 100 different elements alone)
- Authors of HTML documents should not use elements, attributes, or attribute values for purposes other than their intended semantic purpose
  $\rightsquigarrow$ otherwise documents might not be processed correctly
  (still, most authors violate this rule)
- HTML5 follows the separation of concerns design principle:
  a system should be divided into parts with functionality that overlaps as little as possible
  $\rightsquigarrow$ in HTML5 semantics and presentation are (mostly) separated
- For the full specification of the most recent version see

  S. Faulkner, A. Eicholz, T. Leithead, A. Danilo, S. Moon, editors:
  HTML 5.2. W3C Recommendation, 14 December 2017.
  `https://www.w3.org/TR/html52/` (accessed 09 September 2019)

Elements, Attributes and Values

- Most elements consist of a start tag and a matching end tag, with some content in between
- The general form of a start tag

  <*tagName* *attrib1*="*value1*" ... *attribN*="*valueN*">

  where *tagName* is a non-empty sequence of alphanumeric ASCII chars,
  attrib1,...,*attribN*, $N \geq 0$, are attributes and
  value1,...,*valueN*, $N \geq 0$, are attribute values
- A end tag / closing tag takes the form

  </*tagName*>

Examples:

```
<title>My first HTML document</title>
<a href="http://cgi.csc.liv.ac.uk/">CS Website</a>
```

Elements, Attributes and Values

- So-called void elements only have a start tag

  `area` `base` `br` `col` `embed` `hr` `img`
  `input` `keygen` `link` `meta` `param` `source` `track` `wbr`

- The start tags of void elements can be made self-closing by ending the tag with /> instead of >, optionally preceded by a space

  Examples:

```
<br>        <br/>        <br />
<img alt="Picture of Crowne Plaza" src="pic.png">
<img alt="Picture of Crowne Plaza" src="pic.png"/>
<img alt="Picture of Crowne Plaza" src="pic.png" />
```

- Comments take the form

```
<!-- Comment -->
```

  and cannot be nested

Elements, Attributes and Values

- HTML5 distinguished between different categories of attributes
  - Required attributes: needed by elements of a particular type to function correctly
  - Optional attributes: used to modify the default functionality of an element
  - Standard attributes: supported by a large number of element types
  - Event attributes: used to link an element to code that is run if a particular event happens in the element's context
- Standard attributes include:
  - `id`: meant to provide a document-wide unique identifier for an element that can be used to refer to that specifc element
  - `class`: assigns an element to a named group either for semantic or for presentation purposes
  - `title`: assigns a subtextual explanation to an element; in a web browser typically shown if the mouse 'hovers' over the element
  - `style`: allows to change the presentation of an element

Non-ASCII Characters

- The HTML5 specification defines a large number of named characters with the general format `&name;`
  ⇝ allows access to non-ASCII and reserved characters

Examples	Named char	Rendered as
	`&acirc;`	â
	`&auml;`	ä
	`&alpha;`	α

Named char	Rendered as
`<`	<
`>`	>
`&`	&

- Arbitrary characters can also be accessed using `&#dec;` and `&#xhex;` where *dec* and *hex* are decimal and hexadecimal encodings for a character

Examples	Named char	Rendered as
	`&#x000E2;`	â
	`&#x000E4;`	ä
	`&#x003B1;`	α

Named char	Rendered as
`&#x0003C;`	<
`&#x0003E;`	>
`&#x00026;`	&

Revision and Further Reading

Read

- Chapter 2: How the Web Works
- Chapter 4: Creating a Simple Web Page

of

J. Niederst Robbins: Learning Web Design: A Beginner's Guide to HTML, CSS, JavaScript, and Web Graphics (5th ed).
O'Reilly, 2018.
E-book `https://library.liv.ac.uk/record=b5647021`