

Kapitel II.4: HTML - Das System hinter dem Web¹

Dr. Frank Krüger

Fachbereich 23, Johannes Gutenberg-Universität Mainz

e-mail: krueger@acm.org

Gliederung

1. Was ist HTML, die [Hypertext Markup Language](#)?
2. [HTML im Überblick \(Tutorial\)](#)

siehe auch:

- Anhang C0: HTML zum [Nachschlagen](#)
 - [Befehle](#)
 - Darstellung von [Sonderzeichen](#)
- [Anhang C1: Konvertierprogramm RTFtoHTML 3.x](#) (vgl. auch offizielle Doku)
- Anhang C2:
Editorfunktionen von NETSCAPE NAVIGATOR 3.0 GOLD ([Macintosh](#) bzw. [Windows](#);
vgl. auch [offizielle Dokumentation](#)) und AOLPRESS (aus: [offizieller Dokumentation](#))
- Anhang C3 (In Vorbereitung):
Editorfunktionen der MICROSOFT INTERNET-ASSISTENTEN FÜR WORD, EXCEL UND
POWERPOINT

1. Was ist HTML?

A. spezielle Literatur:

- Rieger, (1994): SGML für die Praxis, Kap. 12.2, S. 254-295
- iX-Artikelserie seit 1994 (siehe entsprechende Abschnitte)

B. elektronische Anleitungen:

- NCSA Primer (Beginner's Guide to HTML)
- HTML: Specification of the Document Type Definition (lokale Kopie)

HTML (Hypertext Markup Language) ist eine einheitliche, formale Beschreibungssprache für die sog. Web-Seiten als Grundelemente des World-Wide Web. Mit ihr ist es u.a. möglich, Verweise zwischen diesen Seiten interaktiv darzustellen (sog. Hypertext).

Grundlage: ISO-Standard **SGML** (Standard Generalized Markup Language) zur Darstellung und zum Austausch komplexer elektronischer Dokumente z.B. umfangreicher technischer Handbücher, HTML ist allerdings eine stark vereinfachte Form einer sog. **DTD** (Document Type Definition) und entspricht in vielen Teilen nicht dem Standard.

Prinzip: Grundelement sind die sog. **Tags**, also Markierungen (z.B. `<h1>`) als Metainformation zur Verarbeitung der Dokumente.

Damit werden die Textteile

1. nach ihrer **logischen** Rolle (z.B. eine Überschrift als `<h1>Text</h1>`) gekennzeichnet und nicht - wie in PC-Textverarbeitungssystemen, etwa MS Word üblich - physikalisch (z.B. *Times 14 Pkt, Fett*) formatiert. Die tatsächliche Erscheinungsweise wird erst beim Benutzer durch den "Browser" (das Programm, das das Dokument liest) festgelegt. So wird z.B. eine Überschrift in einem graphischen Browser zwar in einer fetten und etwas größeren Schrift wiedergegeben, ein akustisches System könnte diese Hervorhebung aber durch eine stärkere Betonung und längere Pausen kenntlich machen;
2. universell austauschbar, da nur die Standard-ISO-Zeichensätze (oder andere genormte Zeichensätze, etwa für Japanisch) verwendet werden. Auch nationale und mathematische Sonderzeichen wie Umlaute oder das griechische Alphabet werden umschrieben (siehe [Abschnitt 3.2, Entitäten](#)).

¹: Eine elektronische Version dieser Unterlagen findet sich im WWW unter <http://www.fask.uni-mainz.de/cafl/kurse/komm/komm-24.html> (Kleinschreibung beachten!)

Beispiel:

Der so ausgezeichnete Text	erscheint - etwa in Netscape - als:
<code><h1>&Uuml;berschrift</h1><p>Das ist ein Beispielsabsatz mit einem Verweisauf die Seite wohin.html</code>	Überschrift Das ist ein Beispielsabsatz mit einem Verweis auf die Seite wohin.html

Dazu gilt:

- In der Regel bestehen die Markierungen aus einem Anfang- (`<h1>`) und Ende-Tag (`</h1>`), wobei Groß-/Kleinschreibung hier (ausnahmsweise) keine Rolle spielt
- Sonderzeichen (hier das Ü) werden durch sog. Escape-Sequenzen (`&+Grundzeichen+Art+;`) umschrieben, weitere Zeichen vgl. die Liste der sog. Entitäten, etwa für westeuropäische Sprachen ([ISO 8859-1](#))
- Grundsätzlich muß jeder Absatz mit einem Tag für Paragraph (`<p>`) oder Zeilenumbruch (`
`) **begonnen** werden..
- Zu dem Hypertext-Verweis (sog. Anker) siehe Hypertextfunktionen

2. HTML im Überblick - Ein Tutorial

Im folgenden soll nun der (derzeitige) Befehlsumfang von HTML dargestellt werden. Dabei wird jeweils sowohl auf die (nach SGML) korrekte Form, als auch die von vielen Browsern (z.B. Netscape) zugelassenen Vereinfachung bzw. tolerierten Abweichungen hingewiesen. Außerdem sind einige Erweiterungen des bei uns benutzten Browsers Netscape aufgeführt, die aber teilweise bei anderen Programmen zu Darstellungsproblemen führen können!

Alle Textdokumente bestehen aus dem **Head**, der allgemeine Meta-Information zum Dokument enthält und dem **Body** mit dem eigentlichen Dokumentinhalt. Andere Dokumenttypen (z.B. Gopher, WAIS, News mit etwas anders strukturiertem Textinhalt oder die Grafikformate GIF und JPEG) werden beim Einlesen entsprechend interpretiert und umgesetzt. Eine Besonderheit sind die mit Netscape über die Funktion Bookmark gesammelten Lesezeichen, die in eine eigene Datei exportiert werden können und so einen eigenen HTML-Dateityp mit speziellen Markierungen bilden.

• Darstellungskonventionen

- `GROß` = in KAPITÄLCHEN genauer Wortlaut eines Befehls bzw. dessen Attribute. Die Groß-/Kleinschreibung ist jedoch für die Interpretation der Anweisungen egal, ganz im Gegensatz zu
- `{Text}` = in Groß-/Kleinschreibung einsetzbarer Text, in vielen Fällen ist hier die Schreibweise bedeutungstragend! Umlaute bitte vermeiden
- `#` = einsetzbarer Zahlenwert

- `[]` = optionale Bestandteile
- `|` = trennt Alternativen, die sich in der Regel gegenseitig ausschließen
- Ende-Tags der Form `</TAG>` sind immer dann zwingend notwendig, wenn sie bei dem entsprechenden Element mit aufgeführt werden.

A. Head

Der **Head** enthält als Kopfteil des Dokumentes alle übergreifenden Angaben und steuert ggf. auch die gesamte Wiedergabe des Textes.

Die wichtigsten Elemente:

- `<TITLE>{Titel des Dokumentes}</TITLE>` erscheint als Fenstername unter der Menüzeile
Im Unterschied zur obersten Überschriftenebene (h1) sollte der Titel etwas allgemeiner gehalten sein und bei einem längeren Text sowohl Haupt- als auch Kapitelüberschrift enthalten. Andererseits wird er beim Abspeichern einer Seite etwa in Netscape als Dateiname vorgeschlagen und sollte deshalb für DOS-/Windows-Benutzer nicht länger als 8 Zeichen sein.
- `<META>` ist besonders wichtig geworden, seitdem zahlreiche automatische Suchsysteme (sog. Roboter) die hier enthaltenen Parameter für ihre Indexierung auswerten. Denn mit `META HTTP EQUIV="Keywords" CONTENT="{text text text}` können z.B. Stichwörter für diese Indexierung vorgegeben werden.
- `<BASE HREF="{url}">` erzeugt einen feste Referenzadresse für das aktuelle Dokument
Hier bitte nicht verwenden, da Dokumente in Mainz gespiegelt (d.h. dorthin kopiert und von dort nach außen angeboten) werden.

Weitere Elemente, die wegen ihrer sehr allgemeinen Bedeutung sowohl im HEAD als auch BODY erscheinen können:

- `<ISINDEX [PROMPT="{erläuterung}"]>` erzeugt ein Eingabefeld für eine Stichwortsuche in dem aktuellen Dokument. Das Attribut PROMPT ist ein Erweiterung von Netscape und wird von anderen Browsern ignoriert.
- `<LINK REV="made" HREF=MAILTO:{name@adresse}>` ist die beste Art, ein Dokument zu signieren, da die Urheberschaft mit der E-mail-Adresse gekennzeichnet wird und so Rückmeldungen direkt per e-mail geschickt werden können.
- `<!--z.B. Datum und Version-->` Kennzeichnung von Kommentaren, kann auch verwendet werden, um Markup zeit- bzw. probeweise auszublenden.



Tip: Der gesamte Head kann weitgehend unverändert aus der Dokumentvorlage [WWW_FASK.dot](#) (für Word und Konvertierung) bzw. dem Referenzdokument [WW_HTML.html](#) (für eine direkte Bearbeitung mit Editoren) übernommen werden (Großschreibung und 8.3-Namenskonvention wegen DOS-Kompatibilität!)

B. allgemeine Textdarstellung

Besonders wichtig sind die Absatzelemente:

- `<P>` und `
` wie oben im Beispiel erwähnt. Der Unterschied liegt bei der Darstellung von geringerem Textabstand mit `
`. Beide können auch verschiedene Parameter enthalten, z.B. `<P ALIGN={CENTER}>` für zentrierten Text und bei Netscape `<BR CLEAR={LEFT|RIGHT}>` um den Text unterhalb einer Grafik fortzusetzen.
- `<H1>`, `<H2>` bis `<H6>Text</H#>` markieren die sechs Überschriftenebenen, wobei die Hierarchie streng eingehalten werden sollte, also keine Ebene übersprungen werden darf.
- `<HR>` fügt einen horizontalen Strich (horizontal ruler) ein, wobei Netscape einige zusätzliche Parameter erlaubt (vgl. Referenzteil).

C. Hypertextfunktionen (Verweise)

Grundlegendes Element ist

- `{verweistext}` als ein Verweisanke. Der Verweistext erscheint farbig und unterstrichen (wie im obigen Beispiel). URL (Uniform Resource Locator) ist die absolute oder relative Adresse von dem Dokument auf das verwiesen wird (vgl. [Kapitel I.4. Adressen und MIME-Datentypen](#)).

Bei den Adressen spielt aufgrund der Eigenschaften des UNIX-Betriebssystems die Groß-/Kleinschreibung eine unterscheidende Rolle: *Welcome.html* ist nicht gleich *welcome.html*.



- `` definiert den Namen eines Dokumententeiles, wie er in der URL als *#Fragment* angesprochen werden kann. Dies ist v.a. zur Gliederung von längeren Dokumenten sinnvoll, und um von einer Übersicht am Dokumentanfang direkt zu den einzelnen Abschnitten zu gelangen.

D. logische Auszeichnungen von Zeichen und Absätzen

Am interessantesten für sprach- und kulturwissenschaftliche Texte dürften folgende Markierungen sein (zur tatsächlichen Formatierung z.B. in Netscape, vgl. die Kommentare im [Referenzteil](#)):

- `<CITE>{Zitat}</CITE>` und `<BLOCKQUOTE>{eingerrückter Absatz, z.B. Zitat}</BLOCKQUOTE>` für zitierte Teilsätze bzw. ganze Absätze
- `<DFN>{Definition}</DFN>` für Definitionen, die dann automatisch in ein Glossar übernommen werden können und dort als `<DD>`, Definitionstext erscheinen
- `{Emphasis}` und `{starke Hervorhebung}`
- `<SAMP>{Beispielstext}</SAMP>`, um ein Beispiel herauszuheben

- `<ADDRESS>{vollständige postalische und elektronische Kontaktadresse}</ADDRESS>`, wie sie auch bei Aufsätzen angegeben wird

Weitere Auszeichnungen sind eher naturwissenschaftlich-technischer Art und hier daher oft mit abweichenden Bedeutungen belegt (vgl. [für den Server des FASK](#)):

- `<CODE>{Programmcode}</CODE>`
- `<KBD>{Tastatureingabe des Benutzers}</KBD>`
- `<VAR>{Variablen oder Parameter z.B. in Formeln oder Algorithmen}</VAR>`

E. physikalische Auszeichnungen

Nach einer strengen Auffassung des SGML-Standards dürften diese Auszeichnungsformate gar nicht vorkommen, sind aber aus praktischen Gründen implementiert worden. Sie sind trotzdem nach Möglichkeit zu vermeiden.

- `<TT>{Schreibmaschinenschrift (Teletype) mit festen Abständen}</TT>` für einzelne Zeichen und `<PRE>{Fest formatierte Absätze (Preformatted text)}</PRE>` z.B. für Tabellen.
In der nächsten Version von HTML (3.2) sollen dazu entsprechende logische Tabellen-Tags eingeführt werden.
- `{Fett (Bold)}`, `<I>{Kursiv (Italic)}</I>`
- `^{Hochgestellt (superior)}`, etwa Fußnotenzeichen

NETSCAPE bietet hierzu außerdem folgende Erweiterungen

- `<BASEFONT SIZE = #>`, wobei # auf 3 voreingestellt ist und die Werte 1-7 annehmen kann. Davon ausgehend kann die Schriftgröße verändert werden:
- `` und zwar relativ (+3,-2) oder absolut (bei einem `BASEFONT=3` ergeben die Beispiele die Schriftgröße 6 bzw. 1)

F. Listen und Aufzählungen

Zusätzlich zu den Überschriften gibt es noch drei verschiedene Arten von Listen und Aufzählungen (früher 5, aber MENU und DIR sind inzwischen entfallen):

- `{Text1}{Text2}` erzeugt eine numerierte Liste (Ordered List), wobei `` (vor jedem neuen Eintrag!) einen Zeilenumbruch schon beinhaltet!.
- `{Text1}{Text2}` erzeugt analog zu oben eine Liste mit Aufzählungszeichen.

Beide Listentypen sind sowohl innerhalb des gleichen Types als auch mit dem anderen auf jeweils drei Ebenen schachtelbar. Dabei darf keine Ebene übersprungen werden. Die beiden Unterebenen werden durchnummeriert:

Beispiel: `{Text}<UL1>{Text} </UL1>`

Eine ähnliche Struktur haben die glossarähnlichen Definitionslisten, die aber nicht schachtelbar sind:

- `<DL>` leitet die Definitionsliste ein bzw. schließt sie ab `</DL>`
- `<DD>{Definitonstext}` entspricht dem `` der anderen Listenformate. Beide Teile können aus mehreren Absätzen bestehen.

Beispiel: `<DL><DT>{Definitionseintrag, z.B. Term}</DT>
<DD>{Definitionstext, z.B. Erläuterung}</DD></DL>`

G. Inline-Grafiken und interaktive Karten

Literatur:

- Klute, R.: Sensitive Bilder, iX 10/1994, S. 190-192

Grafiken und Bilder können sowohl in einem eigenen Fenster (mit dem Hypertext-Verweis ``) als auch zusammen mit Text als sog. **inline-Grafik** dargestellt werden. Dazu gelten folgende Auszeichnungsvorschriften:

- `` bezeichnet eine Grafik NAME, die im aktuellen Text dargestellt wird. Kann keine Grafik dargestellt werden, wird alternativ der TEXT angezeigt.
Zugelassene Grafikformate sind GIF (v.a. für Zeichnungen geeignet) und JPG (für Fotos).



Tip: Aus Gründen der Netzwerkökonomie ist es empfehlenswert, größere Grafiken im Text nur als kleines Ikon darzustellen, und mit der vollen Größe als Verweis zu verbinden (siehe oben)

Netscape bietet hier als zusätzliche Funktionalität die Möglichkeit, sowohl die Größe als auch die Ausrichtung der Grafik anzugeben, wobei die einzelnen Erweiterungen bedeuten:

- `ALIGN` läßt eine genaue Bestimmung zu, wie sich die Grafik zu dem Text und anderen Objekten verhalten soll:
- `WIDTH=# HEIGHT=#` beschleunigen das Laden der Bilder, da der Browser die Größe nicht selbst berechnen muß
- `VSPACE=# HSPACE=#` kontrollieren den Abstand des Bildes nach oben und unten bzw. links und rechts (v.a. für schwimmende Bilder)

Grafiken können aber nicht nur zur Präsentation von Information verwendet werden, sondern als sensitive, d.h. "anklickbare" Karten auch zur Markierung von **Bildteilen** als Verweise eingesetzt werden.

Mit dem Attribut `ISMAP` im Befehle `` wird ein Bild als sensitive Karte kennzeichnet und beim Anklicken werden die jeweiligen Mauskoordinaten an das Programm `HTIMAGE` auf dem `WWW-SERVER` weitergeleitet und ausgewertet. Abhängig von den für dieses Bild gespeicherten sensitiven Bereichen - geometrische Figuren wie Rechtecke oder Kreise - wird dann der Verweis zu dem Bereich aktiviert, in dem der Mausklick erfolgte.

Da diese Methode aber die relativ komplizierte Installation von Zusatzprogrammen auf dem Server erfordert, hat Netscape eine "lokale" Variante eingeführt, die ohne serverseitige Programminstallationen auskommt:

Das Attribut `USEMAP` erfüllt die gleiche Aufgabe wie `ISMAP`, erlaubt aber eine Definition der Karte im gleichen (oder einem anderen) HTML-Dokument. Diese Definition der sensitive Bereiche erfolgt mit dem neu eingeführten Tags:

```
<MAP NAME="{name}">
<AREA [SHAPE="rect|circle|poly|default"] COORDS=#,#,#[,#] [HREF=URL]
[NOHREF]>
</MAP>
```

Falls keine Form (SHAPE) vorgegeben ist, wird ein Rechteck erzeugt. Der Parameter `COORDS` bestimmt die Ausmaße der Figur in Pixel² und in der Reihenfolge:

- *links,oben,rechts,unten* bei Rechtecken
- *x (horizontal), y (vertikal) (=Mittelpunkt),Radius* bei Kreisen
- *x,y x,y [x,y]...* bei Vielecken (Polygonen)
- ohne Angabe von Koordinaten bei Default, dem Standardverweis für alle Bereiche eines Bildes, für das keine anderen Zieladressen definiert sind.

H. Tabellen

Mit Tabellen können Texte und Grafiken in einem HTML-Dokument übersichtlich angeordnet werden. Mit den folgenden Befehlen lassen sich als Tabellen erkennbare oder lediglich layouttechnisch besonders angeordnete Dokumentteile (z.B. als rechtsbündiger Text) erstellen.

Die wichtigsten Befehle hierzu sind:

- `<TABLE></TABLE>` definiert eine Tabelle, wobei verschiedene Parameter gesetzt werden können:
 - `<TABLE BORDER=#>` erzeugt einen Rand um die Tabellen und alle sog. Zellen (Kreuzpunkte von Spalten und Zeilen)
 - `<TABLE WIDTH=#|%>` legt die Breite der Tabelle in absoluten Pixeln oder als Prozentsatz der Fensterbreite fest.
 - weitere Optionen sind `CELLSPACING=#` und `CELLPADDING=#`

² : Die Pixel-Koordinaten werden z.B. in professionellen Grafikprogrammen (z.B. Photoshop) angezeigt, können aber auch durch eine Nachbearbeitung der Grafik z.B. in WebMap (Macintosh) durch direkte Manipulation ermittelt und abgespeichert werden. Für die Konvertierung der Daten in das `USEMAP`-Format steht ein Makro zur Verfügung.

- Hauptbestandteile der Tabelle sind die Zeilen (<TR></TR>), die jeweils wiederum aus mehreren "Daten-"Zellen <TD></TD> bzw. Beschriftungen (<TH></TH>, fett und zentriert dargestellt) bestehen.
- Die jeweiligen Optionen für diese Elemente sind:
 - <TR|TH|TD ALIGN=LEFT|RIGHT|CENTER VALIGN=TOP|MIDDLE|BOTTOM> für die horizontale und vertikale Ausrichtung des Text bzw. Objekte innerhalb der Zeilen und Zellen.
- Für Tabellenbeschriftungen und -daten lassen sich zusätzlich als Formatierungseigenschaften angeben:
 - <TH|TD WIDTH=?|% COLSPAN=? ROWSPAN=? NOWRAP>, d.h. die absolute bzw. relative Breite und ob sich eine Zelle über mehrere Spalten bzw. Zeilen erstrecken soll.
- Außerdem kann eine Gesamtüberschrift für die Tabelle definiert werden:
 - <CAPTION ALIGN=TOP|BOTTOM></CAPTION>

I. Formularelemente

Formulare (neu mit HTML 2.0 eingeführt) ermöglichen die Rückmeldung von Information durch den Benutzer und erweitern dadurch wesentlich die Interaktionsmöglichkeiten.

Beispiele für Anwendungen sind:

- *Rückmeldung für weitere Informationen oder Kommentare an den Anbieter einer Seite*
- *Auswahl und Eingabe von Kriterien für die Suche in einer Datenbank*

Allerdings ist die Weiterverarbeitung dieser Rückmeldungen recht komplex, da dazu im allgemeinen in einer Programmiersprache geschriebene Routinen erforderlich sind, mit denen der Server über das sog. CGI (= Common Gateway Interface) kommuniziert.

Dennoch sollen hier der Vollständigkeit halber die Elemente aufgeführt werden.

- <FORM [METHOD=GET|PUT]>{Formularinhalt inkl. weiterer Elemente}</FORM> definiert ein Formular, wobei GET und PUT zwei unterschiedliche Zugriffsmethoden sind. Als Inhalt sind alle anderen Tags erlaubt plus
- <INPUT NAME="{text}"> mit folgenden Attributen und Parameterwerten (Achtung: Kleinschreibung beachten!):
 - TYPE ="text|password|" MAXLENGTH=# SIZE=# definieren Textfeld mit einer maximalen Länge MAXLENGTH und der angezeigten Größe SIZE (ist SIZE kleiner, kann das Feld gescrollt werden)

- TYPE="checkbox|radio" CHECKED unterscheiden dadurch, daß bei CHECKBOX mehrere Werte ausgewählt werden können und bei RADIO immer nur einer.
- TYPE="image" SRC="{name}" ALIGN="top|bottom|middle" stellt ein sensitives Bild (s.o. unter f) dar, wobei es aber nicht so einfach zu handhaben ist wie der ISMAP-Parameter
- TYPE="hidden" zeigt kein Feld an und dient zur Speicherung von Werten vorheriger Übertragungen zum Server
- TYPE="reset|submit" sind zwei Standardtasten, wobei RESET alle Werte auf die Voreinstellungen zurücksetzt und SUBMIT zur Übertragung des Formularinhaltes an den Server angeklickt wird
- VALUE="{text}" gibt bei Textfeldern den Initialwert des Feldes an und bei Checkboxes bzw. Radio button den zurückgemeldeten Attributwert (wird nicht angezeigt, oft identisch mit Namen).
- <TEXTAREA NAME="{text}" ROWS=# Cols=#>{Inhalt}</TEXTAREA> erzeugt mehrzeilige Eingabefelder mit INHALT als Ausgangstext
- <SELECT NAME="{name}" SIZE=# [MULTIPLE] ...</SELECT> läßt eine kompaktere Darstellung von Checkboxes oder Buttons zu, als es bei INPUT der Fall wäre. SIZE gibt an, wieviel Elemente in dem aufklappbaren Menü sichtbar sein sollen und MULTIPLE, ob mehrere Werte ausgewählt werden können. Diese Werte werden innerhalb der SELECT-Markierung spezifiziert mit
- <OPTION [SELECTED] VALUE="{wert}">{Beschreibung}, wobei SELECTED den Eintrag als ausgewählt angibt, VALUE den zu übergebenden Wert bei einer Auswahl und Beschreibung unabhängig davon ein erläuternder Text.