Proseminar“Intelligente IR in WWW - SoSe 2006

Darstellung von Dokumenten in WWW -HTML und CSS -

Cristina Vertan

Informationserhebung und -filterung

- Das WWW kann man als riesige Datenquelle auffassen, die aber ohne Filterung meist nutzlos ist. Beispiel:
- Eine Anfrage *„Computer + Schule“* brachte
  - bei Fireball 2.091.105 Fundstellen,
  - bei Yahoo 100 Fundstellen und 14 Nachrichten, darunter auch die Taifun-Nachricht:

    „Der Taifun verursachte Stromausfälle sowie Störungen im Luft- und Straßenverkehr. Mehr als 250.000 Haushalte waren ohne Strom. Die Finanzmärkte in der Hauptstadt Taipeh blieben am Montag geschlossen, ebenso Regierungsbüros und Schulen. Die Arbeit im Industriegebiet im Norden Taiwans wurde nicht ausgesetzt. Auch das Zentrum der Computerchip-Produzenten von Hsinchu war nicht betroffen.“

Konsequenz

- Die Internet-Texte müssen aufbereitet werden, um die richtige Information und genau diese zu finden. Was wir finden, ist einerseits zu viel und andererseits zu wenig:
- Sei
  - a die Menge der relevanten Treffer einer Suche
  - b die Menge der nicht relevanten Fundstellen und
  - c die Menge der nicht gefundenen relevanten Daten
  - So ist der Recall (Vollständigkeit) = $\frac{a}{a+c}$ und
- Die Precision (Genauigkeit) = $\frac{a}{a+b}$
- Leider ist dieses Maß nur bei Datenbanken wirksam anwendbar, denn nur dort kennt man den Wert von c. Im Internet brauchen wir daher nicht numerische, sondern intelligente Methoden, um überhaupt zu bestimmen, was vernünftige Ergebnisse sind.

Was suchen wir überhaupt?

Fakten?

- Wie groß ist die Entfernung vom Mond zur Erde?
- Welches Bruttosozialprodukt hatte Italien im Jahr 1998?

⇒ Faktenretrieval in Datenbanken

Fundtexte?

- Wo kann ich etwas zur Entwicklung der Chipherstellung lesen?
- Allgemeine Texte zur Steinzeit?

⇒ Stichwort- und Kategorienretrieval im Internet

Antworten?

- Was kann ich in der Schule mit dem Internet machen?
- Wo ist der beste Urlaubsort für mich?

⇒ Data-Mining

Welche Möglichkeiten der Suche haben wir?

1. Web-Texte sind in einer einheitlichen Darstellungs-Sprache organisiert (HTML), Sie enthalten den Text + Gliederungsinformation + wenige allgemeine Daten. Wir haben also mehr als nur die Wörter eines Textes, sondern auch dessen Organisationsinformation (z. B. Überschriften,Verfasser, Datum, Sprache, Suchbegriffe)
2. Mit geeigneten Methoden können wir sogar linguistische Beziehungen finden:
   - Flexionsformen derselben Wörter zusammenfassen,
   - Relevante Wörter von grundsätzlich irrelevanten trennen,
   - Domänenrelevante Wörter auswählen,
   - Syntaktische (damit u.U. logische) Abhängigkeiten berechnen,
3. Wir können gefundene Wörter mit einem Begriffsgerüst (einer Ontologie, z.B. im *Semantic Web*) vergleichen, wir können also (vom Fundwort zur Ontologie) Unter- und Oberbegriffe zur weiteren Suche gegeneinander tauschen oder (von der Ontologie zum Fundwort) Teile einer Ontologie im Netz suchen.
4. Erfolgreiche Suche analysieren und weitere Suchprozesse dadurch steuern (Effiziente Suchstrategien lernen)

Hypertexte, die Grundidee von HTML

- Die Texteinheiten im WWW sind fast ausschließlich Hypertexte.
- Als Hypertext bezeichnet man Texte, die in nicht-linearer Anordnung Inhalte präsentieren. So kann jeder Leser die Reihenfolge und die Granularität des konsumierten Inhalts selbst bestimmen.
- Hypertexte eignen sich daher besonders gut
  - für strukturierte Lerninformation,
  - für Texte, die von sehr unterschiedlichen Lesern auf unterschiedliche Art gelesen werden und
  - Zur synthetischen Präsentation von Informationen aus unterschiedlichen Quellen (virtuelle Webseiten)
  - Aber auch als neue Literaturgattung

Kuhlen, Hypertext 1991

weiter

Granularität

- Eigentlich die Korngröße eines Schüttguts,
- metaphorisch aber auch der Grad der Detailliertheit und Strukturiertheit eines Textes und seiner Inhalte

zurück

Nicht-Linearität

Nicht durch die lokale Reihenfolgerelation verbunden

- Linear sind z.B. normale Bücher, die Präsentation normaler Filme, die meisten Instruktionstexte, Hörereignisse, wie Musik oder Hörspiele
- Nichtlinear sind schon immer z.B. Fahrpläne, in denen man Verbindungen heraussuchen kann, die Logik von Kindererzählungen oder Filmen, die ungezielte Wegesuche.

zurück

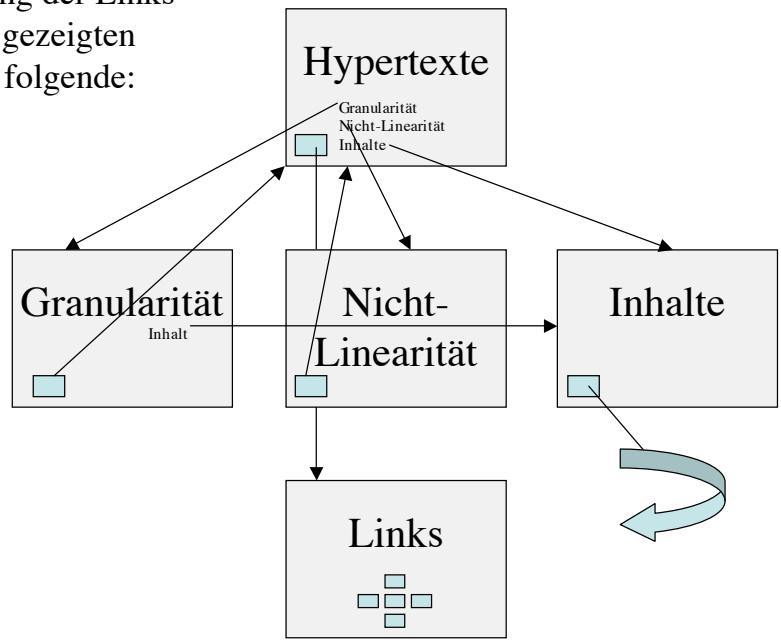
Inhalte und ihre Eigenschaften

- Inhalte in Hypermedien sind meist abgeschlossener als in linearen Medien,
- Sie sind einliniger aber durch die Links eingebettet in unterschiedliche semantische Kontexte.
- Meistens sind sie kleinräumiger (stärker modularisiert) als sequentielle Texte, um ihre mehrfache Einbettung zu erleichtern

zurück

Links

Die Anordnung der Links auf den eben gezeigten Seiten ist die folgende:



Hypertextsysteme

erlauben die einfache Implementierung (Programmierung) von Hypertexten.

- In gewissem Sinne ist PowerPoint auch ein Hypertextsystem, obwohl es nicht vorwiegend für diesen Zweck entworfen ist.
- HTML-Browser, die ihre Funktionalität durch die implementierten „W3C Recommendations“ erhalten. Hier ist das Hypertextsystem eher der Link-Teil des Browsers.
- Reine Hypertextsysteme mit einer Entwicklungsumgebung, wie z.B. das (relativ alte) HyperCard-System von Apple mit seiner Sprache Hypertalk.
- Multimedia-Systeme mit Hypertext-Funktionen, wie Makromedia Director
- Schwierigere Server-Lösungen, wie PHP
- Die Hypertextsprache für WAP-Kommunikation ist WML

HTML

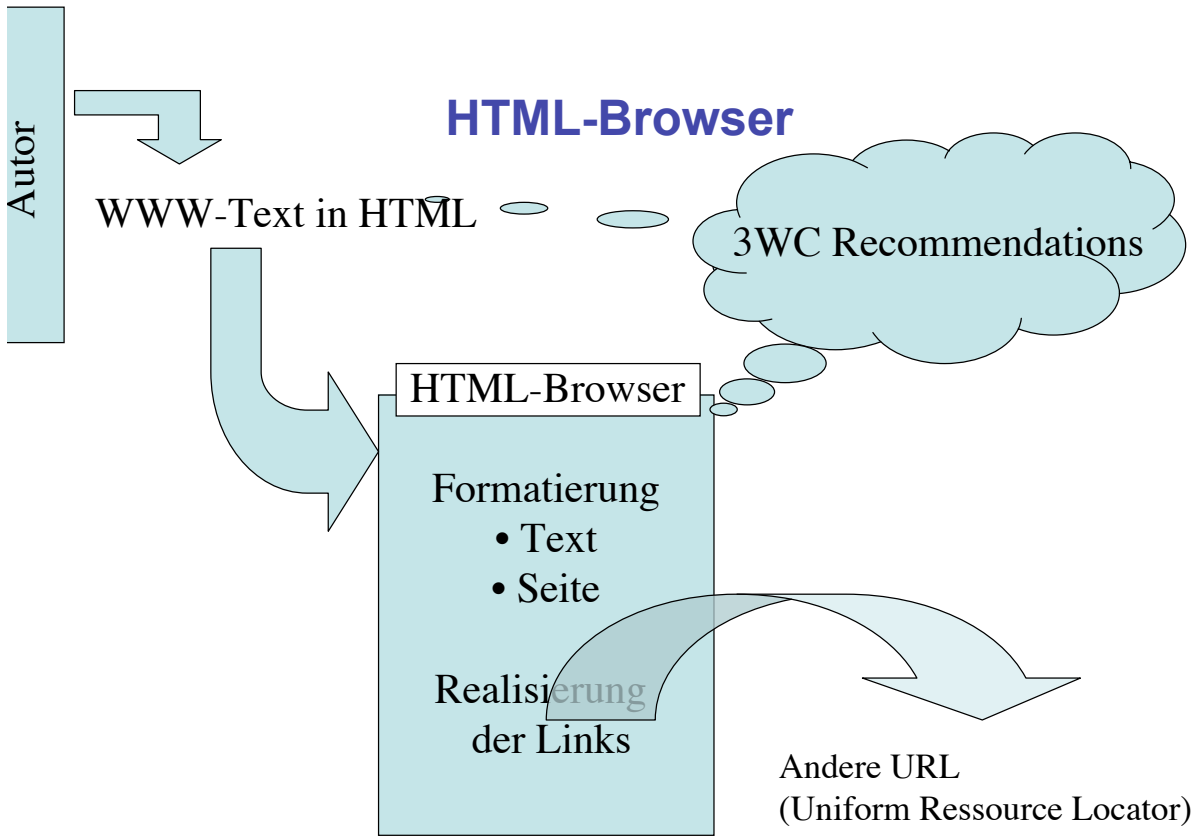
Abkürzung für „Hypertext Markup Language“. Zur Zeit ist die „HTML Specification 4.01“ gültig. Festgelegt ist sie durch die W3C Recommendation vom 24 December 1999.

HTML ist eine Übereinkunft von Web-Entwicklern („W3C“), wie in Web-Browsern plattformunabhängig Webseiten (durch welche Befehle) aussehen sollen.

z.B. soll die Zeichenfolge <hr> immer eine waagerechte Linie erzeugen:

HTML ist aus der Sicht der Benutzer die formale Sprache, in der man Webseiten durch Auszeichnen (Markup) von Texten mit Meta-Zeichenfolgen (Tags) für die empfangenden Computer schreibt.

Eine sehr gute Einführung steht unter: http://de.selfhtml.org/



HTML Texte

- Ein HTML-Text ist ein ASCII-Text, der zusätzliche spezielle Zeichenfolgen zur Darstellung in einem Browser enthält. Ein HTML-Text wird auf allen Plattformen (Betriebssystemen) sehr ähnlich dargestellt. Daher kann man das WWW auch von jedem Rechner aus lesen.
- Die HTML-spezifischen spezielle Zeichenfolgen („Tags“) beschreiben letztlich die logische Struktur einer Webseite, wie etwa:
  - Seitenaufteilung, Überschriften, Absätze, Listen, Tabellen, Zeichenform, Farben, Bildverankerung, und
  - vor allem, die (Hyper-)Links zu anderen Punkten
- Links können ausgehen von Textelementen, graphischen Objekten, Bildern von definierten Regionen einer Seite oder eines Bildes. Sie führen zu andern Absätzen derselben Seite, zu anderen Dokumenten desselben Rechners oder zu Dokumenten anderer weltweiter Rechner (URLs)

Links

Links können ausgehen von

- Textelementen,
- graphischen Objekten,
- Bildern,
- definierten Regionen einer Seite
- definierten Regionen oder eines Bildes.

Sie führen zu

- andern Absätzen derselben Seite,
- anderen Dokumenten desselben Rechners oder
- Dokumenten anderer weltweiter Rechner (URLs, allgemein URIs)

ASCII

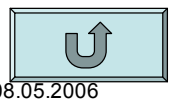
Zeichensatz nach ISO 8859-1, auch „Western“ oder „Latin 1“ genannt.

- Steuerzeichen wie Zeilenwechsel, Leer, Tab, etc.
- Satzzeichen: , . ! ? ; : - " ` ´
- Sonderzeichen § $ % & / \ ( ) < > @ ^ _ [ ] { } | ~
- arithmetische Zeichen: + - * / =
- Zahlen von 0 - 9
- Großbuchstaben: A - Z
- Kleibuchstaben: a - z

In einer zweiten erweiterten Hälfte dieser Tabelle können weitere Zeichen stehen, wie die deutschen Umlaute, das französische ç das dänische å etc.

Alle reservierten Zeichen und die Zeichen der zweiten Hälfte müssen in HTML aber umschrieben werden:

< = < Ä = Ä oder   für Leerzeichen



Die Wirkung von HTML-Tags

HTML:	`<h1>Dies ist eine &Uuml;berschrift </h1>`
Sieht so aus:	Dies ist eine Überschrift

HTML:	`Man kann eine Liste schreiben`
	`<ol>`
	`  <li>mit Zahlen</li>`
	`  <li>mit Spiegelstrichen </li>`
	`  <li>mit "bullets"</li>`
	`</ol>`

Sieht so aus:	Man kann eine Liste schreiben 1. mit Zahlen 2. mit Spiegelstrichen 3. mit "bullets"

HTML von Kopf bis Fuß

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
      "http://www.w3.org/TR/1999/REC-html401-19991224/loose.dtd">
<html lang="de">                                                    Prolog
<head>
   <meta http-equiv="content-type" content="text/html; charset=iso-8859-
   1">
   <title>Eine leere Seite</title>                                  Kopf
   <meta name="generator" content="BBEdit 6.1.1">
</head>
<body>
Hier steht der Text der Web-Seite                                   Körper
</body>
</html>
```

Ausführlicheres Beispiel

```
    Prolog Einträge ....
<html>
<head>  Head-Einträge .....  </head>
<body>
<h1>Vorlesung "Computerphilologie"</h1>
<h2>W.v.Hahn</h2>
<h4>Inhalt:</h4>
<ul>
     <li>01 Intro</li>
     <li>02Theorie</li>
     <li>03Internet</li>
     <li>04Multimedia</li>
     <li>05H&ouml;here Textverarbeitung</li>
     <li>06Lexikalische Repr&auml;sentation</li>
     <li>07Syntaktische Repr&auml;sentation</li>
</ul>
</body>
</html>
```


Das sieht dann so aus:

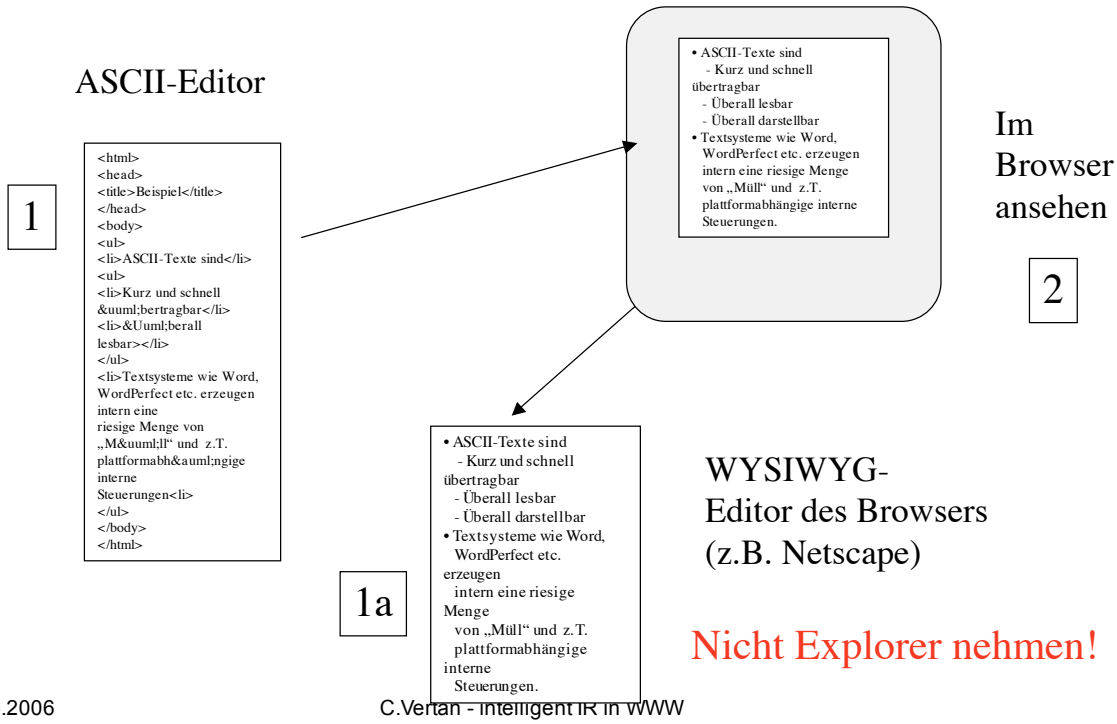
Proseminar "Computerphilologie“

W.v.Hahn

Inhalt:

- Intro
- Theorie
- Internet
- Multimedia
- Höhere Textverarbeitung
- Lexikalische Repräsentation
- Syntaktische Repräsentation

Wie schreibt man HTML-Texte?



Warum nicht einfach ein Textsystem?

ASCII-Texte	*Schreibsystem-Texte*
Kurz, daher schnell übertragbar	Mehr als 100 mal länger durch aufwendige interne Codierung
Mit jedem Editor les- und schreibbar	Gebunden an lizenzierte Software
Plattformunabhängig	Plattformabhängig
Versionsunabhängig	Versionsabhängig

- Nie versuchen: Einen Word-Text (Version 2000) „als HTML speichern“, denn WORD produziert eine entsprechend riesige Menge von HTML-Müll, den man zwar nicht sieht, der aber für Internet-Versendung viel zu sperrig wird.
- Da der MS Internet Explorer keinen Editor (mehr) hat, sondern WORD aufgerufen wird, kann man mit dem Explorer keine effizienten HTML-Texte schreiben oder editieren (schon das öffnen reicht, um den Text aufzublähen).
- Das Lesen von HTML-Texten ist aber bei allen Browsern (ziemlich) ähnlich. Ausnahme z.B.: JavaScript-Interpretation ist im Internet Explorer teilweise anders (dort gibt es JS).

Codierungs-“Müll“

Der Text der vorigen Folie (ohne Tabelle) verbraucht:

Bei Codierung als	*Anzahl Zeichen*	*Anzahl Wörter*
ASCII-Text	429	62
HTML:	1116	179
Word-Code	41.984	5956
WORD → HTML	6255	1002

Was tut der Netscape Composer?

- Man schreibt den Text ungefähr wie mit einem Textsystem.
- Für Formatierung hat man eine Menüauswahl
- Der Composer produziert intern HTML
- Einzelne Markups werden durch Icons angezeigt
- Man kann sich den eigentlichen HTML-Text durch die Funktion „Page Source“ ansehen, aber nicht editieren
- Man kann im Composer jederzeit bestehenden Text editieren
- Im Navigator (dem eigentlichen Browser) sieht man dann den endgültigen Text.

Konsequenz

1. Als Browser für eigene HTML-Versuche am besten nur Netscape benutzen. Ältere Versionen gibt es kostenlos im Netz unter

   http://home.netscape.com/download/archive.html
2. HTML-Texte schreiben
   - in einem einfachen Editor oder
   - im Netscape Composer.
3. Alternativ: Zwar schreiben in einem Schreibsystem dann aber speichern als „Nur Text“, „ASCII“ oder ähnlich und im Editor/Netscape als HTML-Text bearbeiten
4. HTML ruhig ausprobieren für Offline-Texte, z.B. Exzerpte für eine Seminararbeit
5. Es gibt zahllose (teure) spezielle HTML-Editoren, die HTML nach dem WYSIWYG-Prinzip produzieren. Man sollte aber immer den Code verstehen können.

Eine Webseite „verzieren“: DHTML

- Webseiten mit drehenden Objekten, mit Besucherzählern, wandernder Schrift oder allerhand sonstigem Spielkram, aber auch sinnvollen erweiterten Funktionen für die Navigation enthalten Skripts in der Programmiersprache JavaScript. Global nennt man diese Erweiterungen DHTML (Dynamic HTML). Das ist nicht so einfach, es gibt aber Websites, die nützliche Bausteine mit narrensicheren Anleitungen zum herunterladen anbieten:

  http://javascript.internet.com/
- Höchste Ansprüche kann man nur mit Java Applets der Programmier-sprache Java befriedigen. Das ist aber wirklich schwer zu lernen.
- Alle Formulare, Ausfüllfelder, Knöpfe mit der entsprechenden Datenübertragungen etc. muß man mit CGI machen. Das ist Profi-Sache …
- PHP Code wird direkt in HTML-Dateien notiert. Beim Aufruf führt zunächst der serverseitige PHP-Interpreter den Code aus und erzeugt daraus den endgültigen HTML-Code, der schließlich an den Browser gesendet wird.

HTML 4/ CSS

- Seit HTML4 kann man systematisch den Inhalt eines Dokuments und seine logische Gliederung („stylesheet“) trennen. Auf diese Weise kann man
  - existierenden Dokumenten nachträglich ein gleiches Aussehen zuweisen,
  - für weitere Dokumente zentrale Stilvorgaben formulieren, ohne Templates für Dokumente zu schreiben,
  - Inhalte in andere Dokumente und deren Stil übernehmen
  - Die Standard-Darstellung eines Browsers korrigieren
  - Dieses Prinzip heißt CSS („Cascaded Style Sheet“)

3 Arten von Stylesheets

Style information können Sie

1. An Ort und Stelle in einem HTML-Ausdruck schreiben („direktformatieren“)
2. Im Kopf eines Files zentral zusammenfassen
3. In einem getrennten File zitierbar machen

Beispiel zu 1.:

```
<p style="background-color:#808040; color:#D8FD02;
font-family:'Century Schoolbook',serif; font-size:24pt; letter-spacing:3px;
padding:40px; border:double #D8FD02 4px;"
title="Zitat von Francis Picabia">
Unser Kopf ist rund, damit das Denken die Richtung wechseln kann!
</p>
```

Alle Besispiele sind aus: http://de.selfhtml.org

Zentrale Style Syntax

Syntax	<STYLE> ... </STYLE>
Attribute	
	TYPE=ContentType (Mime-Typ und Sprache)
	man kann auch angeben MEDIA=MediaDesc (Medienspezifik)
	TITLE=Text (title of style sheet)
Inhalt	ein stylesheet
Erscheint in	HEAD

TYPE ist meist „text/css“,

MEDIA kann sein: {screen, tty, tv, projection, handheld, print, braille, aural, all}

Man kann also für verschiedene Medien unterschiedliche Stylesheets machen.

Zentrales HTML Stylesheet

Das Beispiel definiert den Stil für die h1 und h2 Überschriften und führt eine neue Klasse „chap“ ein:

```
<head>
<title>CSS Example< /title >
<style type="text/css">
h1 { font-size: x-large; color: red }
h2 { font-size: large; color: blue }
.chap { font-family:Arial,sans-serif; font-size:20pt;
color:blue;
                    border-bottom-style:solid; border-bottom-
width:3px; border-          color:red; margin:0px; margin-
bottom:16px; }
< style >
< head >
<body> ...
<div class="chap">Eine Kapitelüberschrift</div>   ...
```

Externes Stylesheet

Man kann das „link“-Element in <head> benutzen, um auf exsterne Stylesheets zu verweisen:

```
<html>
<head>
<title>Titel der Datei</title>
<link rel="stylesheet" type="text/css" href="formate.css">
<style type="text/css">
<!--
  ... hier sind zusätzlich auch datei-spezifische Formate erlaubt ...
-->
</style>
</head>
<body>
</body>
</html>
```


Beispiel für externes css File

(Keine HTML-Syntax!)

```
#Kapitel {
font-size:300%; text-transform:capitalize; line-height:80pt;
}
#Definition {
  border-width:3px 16px;
  border-style:dashed;
  border-color:red;
  margin-right:10px;
  padding-top:25px;
  padding-bottom:20px;
  padding-left:20px;
  padding-right:20px;
  text-align:justify;
```

Hier werden in einem File die Styles „Kapitel“ und „Definition“ definiert

XML vs. HTML

Was ist XHTML ?

- Re-definition von HTML-tags entspechend der XML-Syntax
- die Dokumentstruktur bleibt dieselbe

```
<html>
   <head>
    <!--   ... Head-Inhalt ....-->
   <title> Dokumenttitel </title>
   <!-- ...  Head-Inhalt ... -->
 </head>
<body>
    <!-- ... Body-Inhalt ... -->
</body>
</html>
```

- Einige Tags haben in XHTML **NICHT** dieselben Namen wie in HTML

XHTML vs. HTML

- Alle Tag- und Attribut-Namen müssen in Kleinbuchstaben geschrieben werden (weil XML kleine und grosse Buchstaben unterscheidet)
- Es gibt keine optionalen Ende-tags. Alle Tags müssen ein Paar sein (also auch z.B. ...** **)
- Alle leeren (textlosen) Tags enthalten wie in XML ein Leeres-Element-Tag)
  <hr /> (statt <hr></hr>)
- es gibt nur ein einziges `head`- und ein einziges `body`- Element, man nur auch nur ein einziges `frameset` -Element einfügen.
- jedes `head`-Element darf nur ein einziges `title` -Element (Tag) enthalten

XHTML vs. HTML (Beispiel)

```
<html>
<head>
<title> Vorlesung CP
    Content</title>
</head>
<body>
<h1> Vorlesung CP </h1>
<hr / >
<h2>Inhalt</h2>
<ul>
   <li>01 Intro </li>
   <li> 02 Theorie </li>
</ul>
</body>
</html>
```

```
<HTML>

<BODY>
<h1> Vorlesung CP </h1>
<hr>
<h2>Inhalt</h2>
<UL>
   <li>01 Intro
  <li> 02 Theorie
</ul>
</body>
</HTML>
```

Viele HTML-Editoren ergänzen HTML zu XHTML

Strictly Conforming XHTML

- Entspricht einem strikten XML-Formalismus:
  - Spezifiziert, dass das Dokument vollständig XML-formatiert ist:
    <?xml version="1.0" charset="iso-8859-1" ?>
  - benennt eine DTD
    <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict// EN"
    "http://www.w3.org/TR/xhtml1/DTD/strict.dtd" >
  - das <html> Element muß ein "xmnls" Attribut enthalten, um zu spezifizieren wo die Elementnamen definiert sind:
    <html xmnls="http://www.w3.org/TR/xhtml1">
    ...
    </html>

Strictly Conforming XHTML - Beispiel

```
<?xml version="1.0" charset="iso-8859-1" ?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0
    Strict// EN"
 "http://www.w3.org/TR/xhtml1/DTD/strict.dtd" >
<html xmnls="http://www.w3.org/TR/xhtml1">
<head>
<title> Vorlesung CP Content</title>
</head>
<body>
.....
</body>
</html>
```

Extraktion von Strukturinformationen aus HTML Dokumenten

- <b> Fett
- <i> Kursiv
- <p> Textabsätze
- <h1><h2><h3> Titeln
- <u> Unterstirchen
- , <ol>, <dl>,... Listendarstellungen
- Tabelle liefern manchmal keine Information über die Struktur, sie sind nur beutzt um mehr Platz zu schaffen

Häufige Formate in WWW

- PDF - sie sind überall lesbar, alerdings man kann die Inhalte nur visualiseren aber nicht weiter bearbeiten
- Word-Dokumente - fast wie PDF Dateien, dazu ist die Visualisierung von der Benutzers Word-Version abhängig
- HTML und immer mehr XML Dokumente
- Um sichtbar für eine Suchmaschine zu sein jede HTML Seite muss indiziert sein. Je mehr relevante Keywords für das Dokument angegeben sind, desto schneller wird er gefunden.

Grenze von HTML

- Mit HTML-Tags, kann man schwierig etwas über die Inhalte sagen.
- Semantische Verknüpfung zwischen Dokumentteile oder zwischen Dokumente ist es nicht möglich
- Deswegen braucht man aussagekräftigere Sprachen und Datenmodellen wir XML und RDF.