

Vorlesung Netzbasierte Informationssysteme
(WS 2004/05)
Darstellung und Mehrsprachigkeit

Robert Tolksdorf
Freie Universität Berlin
Institut für Informatik
Netzbasierte Informationssysteme
mailto:tolk@inf.fu-berlin.de
<http://www.robert-tolksdorf.de>
<http://nbi.inf.fu-berlin.de>

[1] © Robert Tolksdorf, Berlin

Überblick

[2] © Robert Tolksdorf, Berlin

Überblick

- Darstellung von Inhalten
- Mehrsprachigkeit

[3] © Robert Tolksdorf, Berlin

Darstellung von Inhalten

[4] © Robert Tolksdorf, Berlin

Inhalt und Darstellung

- Auszeichnungssprachen markieren
 - logische Struktur (<h1>, <p>, <table>)
 - Darstellung (, <i>, , <big>)
- Trennung von Inhalt und Darstellung ist aber wichtig für
 - Geräteunabhängigkeit von Informationen (Handy vs. PC)
 - Medienunabhängigkeit von Informationen (Grafik vs. Sprache)
 - Sprachunabhängigkeit von Informationen (" vs. „ vs. »)
 - Mehrkanal Veröffentlichungen (WAP und Web)
 - Verarbeitbarkeit von Informationen

[5] © Robert Tolksdorf, Berlin

Cascading Style Sheets CSS

- ... Mechanismus zur separaten Definition von Stileigenschaften für HTML und XML Dateien (Quelle: <http://www.w3.org/Style/CSS/>)
- Cascading Style Sheets, level 1
 - Ziel: Sprache zur Definition des Darstellungsstils von HTML Dokumenten
 - Status: W3C Recommendation 17 Dec 1996, revised 11 Jan 1999
 - Quelle: <http://www.w3.org/TR/REC-CSS1>
- Cascading Style Sheets, level 2
 - Ziel: Sprache zur Definition des Darstellungsstils von HTML und XML Dokumenten für unterschiedliche Medienarten
 - Status: W3C Recommendation 12-May-1998
 - Quelle: <http://www.w3.org/TR/REC-CSS2>
- Cascading Style Sheets, level 3
 - Ziel: Modularisierte und erweiterte Sprache zur Definition des Darstellungsstils von HTML und XML Dokumenten
 - Status: unterschiedlich, erste Recommendations eventuell April 2003
 - Quelle: <http://www.w3.org/Style/CSS/current-work>

[6] © Robert Tolksdorf, Berlin

CSS Grundidee

- Grundidee:
Zu HTML Tags werden definierte Attribute für Darstellungseigenschaften gesetzt
- CSS-Datei getrennt von HTML-Datei gehalten
- Beispiel: Um Überschriften in grosser blauer Schrift darzustellen:

```
h1 {color: blue; font-size: 22pt; }
```
- CSS definiert
 - Rahmensyntax zur Notation
 - Menge von Attributen
 - Menge von Werten
 - Bedeutung
 - Mechanismen zur Anbindung von Stilinformationen an und in HTML Seiten

[7] © Robert Tolksdorf, Berlin

CSS Anbindung an HTML

- Drei Wege der Einbindung in HTML
 - Mit externem Stylesheet über Verweis im <link>-Tag:

```
<link rel="stylesheet" type="text/css" href="http://www.inf.fu-berlin.de/inst/ag-nbi/include/nbi.css">
```
 - Im HTML Dokument mit dem <style>-Tag:

```
<style>
h1 {color: blue; font-size: 22pt; }
</style>
```

Kompatibel für alte Klienten:

```
<style><!--
h1 {color: blue; font-size: 22pt; }
--></style>
```
 - Bei den einzelnen Elementen im style-Attribut:

```
<h1 style="color: red">Rote Überschrift</h1>
```
- Einbindung innerhalb eines CSS
 - @import url(<http://www.inf.fu-berlin.de/inst/ag-nbi/include/colors.css>);

[8] © Robert Tolksdorf, Berlin

CSS Gruppierungen und Vererbung

- In CSS können Angaben gruppiert werden
 - Mehrere Elemente erhalten gleiche Eigenschaften
h1,h2,h3,h4,h5,h6 {color: blue;}
 - Ein Element erhält mehrere Eigenschaften
h1 {color: blue; font-size: 22pt; }
 - Kombination
h1,h2,h3,h4,h5,h6 {color: blue; font-style: italic}
- Eigenschaften werden „vererbt“
 - Entlang der Element Schachtelung nach „unten“
<h1>Das ist <u>wichtig</u></h1>
 - Allgemeinstes Element in HTML: <body>
 - Eigenschaften die dort gesetzt werden, gelten für alle Elemente bei denen nichts anderes deklariert ist
 - Beispiel: Seite komplett in serifenloser Schrift
body {font-family: arial, helvetica, sans-serif; }

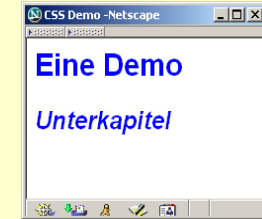
[9] © Robert Tolksdorf, Berlin

Demo

- HTML:

```
<html>
<head>
  <link rel="stylesheet" type="text/css" href="
  "cssdemo.css">
  <title>CSS Demo</title>
</head>
<body>
  <h1>Eine Demo</h1>
  <h2>Unterkapitel</h2>
</body>
```
- CSS:

```
body {font-family: arial, helvetica, sans-serif; }
h1 {color: blue; font-size: 22pt; }
h2,h3,h4,h5,h6 {color: blue; font-style: italic}
body {font-family: arial, helvetica, sans-serif; }
```



[10] © Robert Tolksdorf, Berlin

Wertetypen

- Längen ([+|-]<Zahl><Einheit>)
 - Relative Längen
 - em: Breite des M im aktuellen Font
 - ex: Höhe des x im aktuellen Font
 - px: Referenzpixel auf einem 90 dpi Gerät
 - Absolute Längen
 - in: Zoll (1in=2,54cm)
 - cm: Zentimeter
 - mm: Millimeter (10mm=1cm)
 - pt: Typographischer Punkt (1pt=1/72in)
 - pc: Pica (12pc=1pt)
- Anteilige Größen in Prozent
 - Bei Eigenschaft ist Bezug definiert
 - p { line-height: 120% }: Relativ zum aktuellen font-size
 - h1 { margin-right: 12.3% }: Relativ zum aktuellen margin-right

[11] © Robert Tolksdorf, Berlin

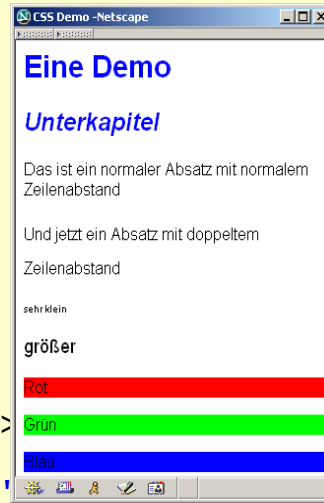
Wertetypen

- Schlüsselwort-Werte
 - Teilweise definiert, z.B. bei font-size
 - absolut: xx-small, x-small, small, medium, large, x-large, xx-large
 - relativ: larger, smaller
- Farbwerte
 - Farbnamen:
aqua, black, blue, fuchsia, gray, green, lime, maroon, navy, olive, purple, red, silver, teal, white, yellow
 - Farbanteile:
 - #rgb color: #F0F
 - #rrggbb color: #FF00FF
 - rgb(val,val,val) color: rgb(255,0,255)
 - rgb(frac,frac,frac) color: rgb(100%,0%,100%)
- URLs
 - background: url(http://www.bg.com/pinkish.gif)

[12] © Robert Tolksdorf, Berlin

Demo

```
<p>Das ist ein normaler Absatz mit normalem Zeilenabstand</p>
<p style="line-height:200%">Und jetzt ein Absatz mit doppeltem Zeilenabstand</p>
<p style="font-size:x-small">sehr klein</p>
<p style="font-size:larger">größer</p>
<p style="background-color:red">Rot</p>
<p style="background-color:#0F0">Grün</p>
<p style="background-color:rgb(0,0,100%)">Blau</p>
```



[13] © Robert Tolksdorf, Berlin

CSS Klassen

- Darstellungseigenschaften für alle Verwendungen eines bestimmten Elements gleich
- Tatsächlich aber vielleicht unterschiedliche Eigenschaften je nach Verwendung → Klassen in CSS
- Deklaration durch Punkt getrennt hinter Element:
h1.largegreen {color: green; font-size: 22pt; }
- Verwendung durch class-Attribut
<h1 class="largegreen">Das ist jetzt anders</h1>
- Deklaration von Eigenschaftsklassen ohne Elementangabe:
.green {color: green;}
- Verwendung bei allen Elementen möglich:
<p class="green">Alles klar</p>

[14] © Robert Tolksdorf, Berlin

HTML Elemente zur reinen Stilbindung

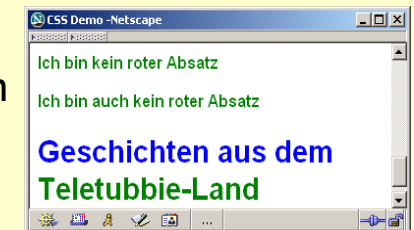
- Zwei Elemente in HTML, mit denen Stil gebunden werden kann ohne dass Darstellung erzeugt wird:
 - <div>: Generisches Block-Element, das andere Blöcke umschließt
 - : Generisches Inline-Element, das im fortlaufenden Text verwendet werden kann
- Attribute: id, lang, dir, title, align, onclick, ondblclick, onmousedown, onmouseup, onmouseover, onmousemove, onmouseout, onkeypress, onkeydown, onkeyup, style, class
- <div class="green"><p>blah blah</p><p>foo foo</p></div>
- <p>Das ist wirklich sehr wichtig.</p>

[15] © Robert Tolksdorf, Berlin

Demo

```
<p class="green"><b>Ich bin kein roter Absatz</b></p>
```

```
<div class="green">
<p><b>Ich bin auch kein roter Absatz</b></p>
</div>
```



```
<h1>Geschichten aus dem
<span class="green">Teletubbie-Land</span></h1>
```

[16] © Robert Tolksdorf, Berlin

CSS Ausnahmen

- Unterschiedliche Eigenschaften je nach Kontext der Verwendung → „contextual selectors“ in CSS
- In Überschrift
`<h1>Das ist wichtig</h1>`
soll Hervorhebung farblich sein:
`h1 em { color: red }`
- Kann verknüpft werden
`ul li { font-size: small }`
`ul ul li { font-size: x-small }`
- Kann auch gemischt werden
`.red h1 {color: blue}`
→ `<h1>` in einem `<div class="red">` ist trotzdem blau

[17] © Robert Tolksdorf, Berlin

CSS Pseudoelemente und Pseudoklassen

- Pseudoklassen (hier beim Anker-Element `<a>`)
 - `a:link`: noch nicht besuchter Zielanker
 - `a:visited`: schon besuchter Zielanker
 - `a:active`: gerade ladender Zielanker
 - `a:hover`: Mauszeiger ist über Element
 - `a:focus`: Element hat Eingabefocus
 - `a:link {text-decoration : none ;}`
`a:hover {text-decoration : underline ;}`
- Typographische Pseudoelement
 - `first-letter` Erster Buchstabe
 - `<p>Guten Tag</p>` wird konzeptionell expandiert zu
`<p><p:firstletter>G</p:firstletter>uten Tag</p>`
 - `first-line` Erste Zeile
 - `p:first-line { font-family: serif }`
 - The first line of an article in Newsweek.
 - `first-child` Erstes Unterelement
- Können mit eigenen Klassen kombiniert werden:
`p:einleitung:first-line { font-family: serif }`

[18] © Robert Tolksdorf, Berlin

Auflösung von Konflikten

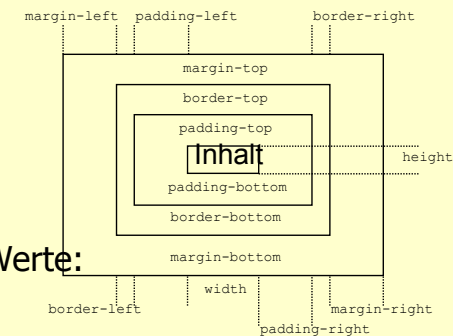
- Darstellungseigenschaften werden an verschiedenen Stellen definiert (Browser, Nutzerpräferenzen, Autor)
- Bei Konflikten gilt folgende ansteigende Gewichtung:
 - Defaults des Darstellers
 - Eventuell persönliches Stylesheet
 - Stylesheet des Dokuments
- Innerhalb von Stylesheets
 - Deklarationen mit Markierung `!important` wichtiger als solche ohne
(`h1 {color: blue !important ; font-size: 22pt; }`)
 - Deklarationen aus höher gewichteter Quelle sind wichtiger als solche aus anderen
 - Spezifischere Deklarationen gewichtiger als allgemeinere
(`h1 em { color: red }` vs. `h1 {color: blue;}`)
 - Textuell spätere Deklaration gewichtiger als frühere (dabei `@import` am Anfang in ihrer Reihenfolge berücksichtigt)

[19] © Robert Tolksdorf, Berlin

Eigenschaften von Inhaltskästen

- Leerraum um Kästen

- `margin-top`: Oben
- `margin-right`: Rechts
- `margin-bottom`: Unten
- `margin-left`: Links
- `margin`: Zusammengefasste Werte:
 - `margin: 1cm`
Auf allen Seiten 1cm
 - `margin: 1cm 2cm`
Oben und unten 1cm, links und rechts 2cm
 - `margin: 1cm 2cm 3cm`
Oben 1cm, rechts 2cm, unten 3cm, links 2cm
 - `margin: 1cm 2cm 3cm 4cm`
Oben 1cm, rechts 2cm, unten 3cm, links 4cm

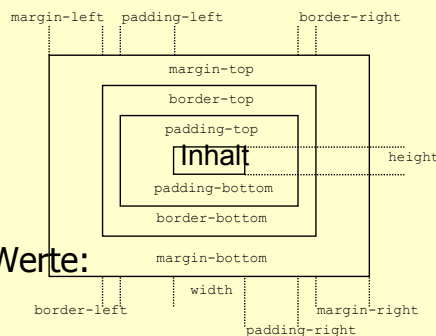


[20] © Robert Tolksdorf, Berlin

Eigenschaften von Inhaltskästen

Leerraum um Inhalt

- padding-top: Oben
- padding-right: Rechts
- padding-bottom: Unten
- padding-left: Links
- padding: Zusammengefasste Werte:
 - padding: 1cm
Auf allen Seiten 1cm
 - padding: 1cm 2cm
Oben und unten 1cm, links und rechts 2cm
 - padding: 1cm 2cm 3cm
Oben 1cm, rechts 2cm, unten 3cm, links 2cm
 - padding: 1cm 2cm 3cm 4cm
Oben 1cm, rechts 2cm, unten 3cm, links 4cm



[21] © Robert Tolksdorf, Berlin

Eigenschaften von Inhaltskästen

- Dicke des Rands um Inhalt + Padding
 - border-top-width: *Dicke Oben*
 - border-right-width: *Dicke Rechts*
 - border-bottom-width: *Dicke Unten*
 - border-left-width: *Dicke Links*
 - border-width: *oben rechts unten links* Zusammengefasste Werte
 - Werte: thin, medium, thick oder Länge
- Farbe des Rands
 - border-color: *oben rechts unten links*
- Linenstil des Rands
 - border-style: *oben rechts unten links*
 - none, dotted, dashed, solid, double, groove, ridge, inset, outset
- Zusammengefasst für Seiten einzeln
 - border-top: *Dicke Linienstil Farbe*
 - Gleiches für border-right, border-bottom, border-left
 - border: *Dicke Linienstil Farbe* für alle vier Seiten gleich

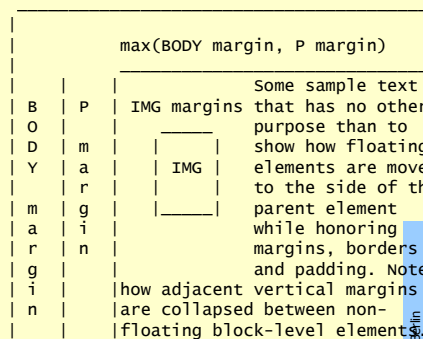
[21] © Robert Tolksdorf, Berlin

Eigenschaften von Inhaltskästen

- width: *Länge* Breite des Inhalts
- height: *Länge* Höhe des Inhalts
- float: *Position* Inhalt kann an den Rand des enthaltenden Elements verschoben werden

The above example could be formatted as:

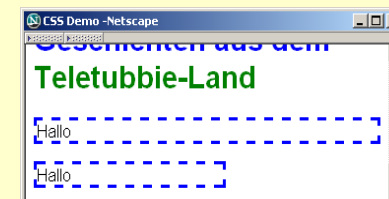
```
<style type="text/css">
img { float: left }
body, p, img { margin: 2em }
</style>
<body>
<p>
Some sample text that has no
other...</p></body>
```



[23] © Robert Tolksdorf, Berlin

Demo

- CSS:
 - .aufwaendig {border: medium dashed blue;}
- HTML:
 - <p class="aufwaendig">Hallo</p>
 - <p class="aufwaendig" style="width:5cm; height=2cm">Hallo</p>



[24] © Robert Tolksdorf, Berlin

Darstellungsklassen und Listen

- `display` Legt Darstellungsklasse eines Elements fest
 - `block`: Erzeugt Darstellungsblock (`<p>...</p>`)
 - `inline`: Wird innerhalb eines Blocks dargestellt (`<i>...</i>`)
 - `list-item`: Ist Listeneintrag (`...`)
 - `none`: Keines der obigen (``)
- Falls `display: list-item` gilt
 - `list-style-type` Nummerierungsart
 - Werte `disc`, `circle`, `square`, `decimal`, `lower-roman`, `upper-roman`, `lower-alpha`, `upper-alpha`, `none`
 - `list-style-image`: *URL* Bild als Markierung
 - `list-style-position` Markierungsposition relativ zum Block
 - Wert: `inside`, `outside`
- `list-style` *Nummerierungsart URL Markierungsposition*
 - Kurznotation

• Blah foo blah
• Blah foo blah

[25] © Robert Tolksdorf, Berlin

Elementfarbe und Hintergrund

- `color`: *Farbe* Farbe des Elements
- `background-color`: *Farbe* Hintergrundfarbe
- `background-image`: *URL* Hintergrundbild
- `background-repeat` Wird das Hintergrundbild wiederholt und wie?
 - Werte: `repeat`, `repeat-x`, `repeat-y`, `no-repeat`
- `background-attachment` Ist Hintergrundbild verankert oder wird es mitgescrollt
 - Werte: `scroll`, `fixed`
- `background-position`
 - Werte: Absolute Länge oder relativ zur Umgebungsgröße
 - Schlüsselworte aus
 - `top`, `center`, `bottom`
 - `left`, `center`, `right`
 - `center right` entspricht 100% 50%

[26] © Robert Tolksdorf, Berlin

Elementfarbe und Hintergrund

- `background` *Farbe Wiederholung Fixierung Position*
 - Kurznotation
- Institutshomepage:

```
body {
  color : #222222;
  background : #ffffff
  url(http://www.inf.fu-berlin.de/styles/inst-title-600x400.jpg)
  no-repeat;
}
```

[27] © Robert Tolksdorf, Berlin

Schrifteigenschaften

- `font-family` Name der Schriftfamilie
 - Generische
 - `serif` (Times)
 - `sans-serif` (Helvetica)
 - `cursive` (*Monotype Corsiva*) (Normale Schrift kursiv: *a*, Kursive Schrift: *a*)
 - `fantasy` (Comic Sans)
 - `monospace` (Courier, Lucida Console)
 - Konkrete Schriftnamen (Arial, Tahoma, Swiss, Garamond)
 - Auswahlversuch in Reihenfolge der Liste

```
body { font-family: "new century schoolbook", serif
}
```
- `font-style` Schriftstil
 - `normal`, *italic*, *oblique*
- `font-variant` Schriftvariante
 - `normal`, `small-caps` (KAPITÄLCHEN)

[28] © Robert Tolksdorf, Berlin

Schrifteigenschaften

- **font-weight** Stärke der Schrift (Schriftzug)
 - Absolute Werte: normal, bold, 100, 200, 300, 400, 500, 600, 700, 800, 900,
 - Relative Werte: bolder, lighter
- **font-size** Größe der Schrift
 - Absolute Werte: xx-small, x-small, small, medium, large, x-large, xx-large, Größenangabe
 - Relative Werte: larger, smaller
- **font *stil* variante [Stärke] Größe Zeilenabstand Schriftfamilie**
 - Kurznotation

[29] © Robert Tolksdorf, Berlin

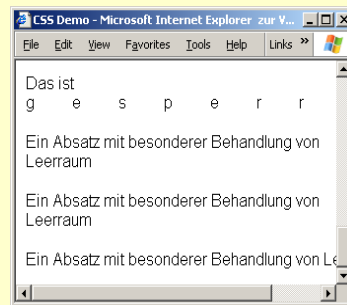
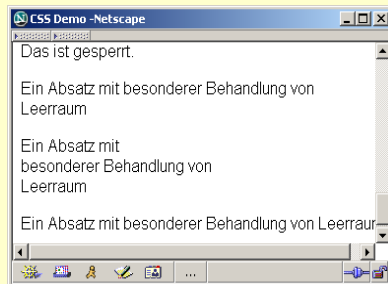
Texteigenschaften

- **word-spacing: Länge**
Zusätzlicher Wortabstand
- **letter-spacing: Länge**
Zusätzlicher Buchstabenabstand (Dicke)
- **white-space** Legt Behandlung von Leerraum fest
 - normal: Leerzeichen fallen zusammen, automatischer Zeilenumbruch (<p>)
 - pre: Leerraum wird beachtet (<pre>...</pre>)
 - nowrap: Kein automatischer Zeilenumbruch mit

[30] © Robert Tolksdorf, Berlin

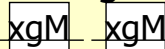
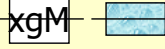
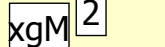
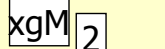


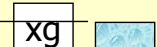

Demo

```
<p>Das ist <span style="letter-spacing=1cm">gesperrt</span>.</p>
<p style="white-space:normal">Ein Absatz mit besonderer Behandlung von Leerraum</p>
<p style="white-space:pre">Ein Absatz mit besonderer Behandlung von Leerraum</p>
<p style="white-space:nowrap">Ein Absatz mit besonderer Behandlung von Leerraum</p>
```



[31] © Robert Tolksdorf, Berlin

Texteigenschaften

- **text-decoration** Textdekoration
 - Werte: none, underline, overline, line-through, blink
- **text-transform** Textveränderung
 - Werte: capitalize, uppercase, lowercase, none
- **vertical-align** Ausrichtung eines Kastens
 - baseline 
 - middle 
 - super 
 - sub 
 - text-top 
 - text-bottom 
 - top 
 - bottom 

[32] © Robert Tolksdorf, Berlin

Texteigenschaften

- `text-align` Horizontale Ausrichtung
 - Werte: `left`, `right`, `center`, `justify`
- `text-indent`: *Länge* Einzug der ersten Zeile
- `line-height` Zeilenabstand
 - Faktor auf Schriftgröße
 - Länge
 - Prozent

CSS2

- Tabellen als Darstellungsobjekt
- Eigenschaften für Tabellen
- Diverse zusätzliche Eigenschaften für Text und Schriften
- ...

CSS2: Medienarten

- Darstellungsstil ist abhängig vom Ausgabemedium
 - Bildschirm
 - Papier
 - Sprache
 - Braille
 - ...
- CSS erlaubt getrennte Stildefinitionen:

```
...
a:link {
  color: #000099;
  text-decoration : none ;
}
@media print {
a:link,a:visited,a:hover,a:active,a:focus {
  text-decoration:none;
  color:blue
}
}
```

Definierte Medienarten

- Medienarten:
 - `all`
 - `aural` Sprachausgabe
 - `braille` Taktile Ausgabe
 - `embossed` Braille Drucker
 - `handheld` Klein, monochrom
 - `print` Drucker
 - `projection` Beamer, Foliendruck
 - `screen` Bildschirm
 - `tty` Textterminals
 - `tv` Geringe Auflösung eingeschränkte Interaktion

Mediengruppen

- Darstellungseigenschaften müssen jeweils von Darstellern für bestimmte Mediengruppen beachtet werden

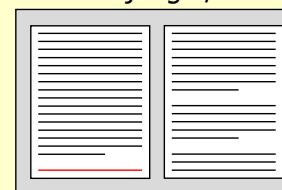
Media Types	Media Groups			
	continuous/paged	visual/aural/tactile	grid/bitmap	interactive/static
aural	continuous	aural	N/A	both
braille	continuous	tactile	grid	both
emboss	paged	tactile	grid	both
handheld	both	visual	both	both
print	paged	visual	bitmap	static
projection	paged	visual	bitmap	static
screen	continuous	visual	bitmap	both
tty	continuous	visual	grid	both
tv	both	visual, aural	bitmap	both

[37] © Robert Tolksdorf, Berlin

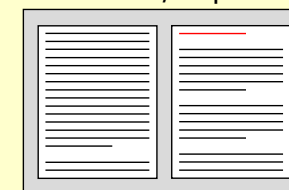
Seitenorientierte Medien

- Zusätzliche Eigenschaften
 - Seite ist umgebender Darstellungskasten
 - page- Eigenschaften
 - Es treten Seitenumbrüche auf
 - Eigenschaften page-break-before, page-break-after, page-break-inside
 - Werte: auto, always, avoid, left, right
 - Eigenschaften orphans, widows
Hurenkinder/Schusterjungen

Schusterjunge / widow



Hurenkind / Orphan



[38] © Robert Tolksdorf, Berlin

Darstellung gesprochener Sprache

- Darstellungseigenschaften betreffen
 - Ton
 - Dauer und Reihenfolge
 - Spracheigenschaften

▪ Beispiel

```
H1, H2, H3, H4, H5, H6 {
  voice-family: paul;
  stress: 20;
  richness: 90;
  cue-before: url("ping.au")
}
P.heidi { azimuth: center-left }
P.peter { azimuth: right }
P.goat { volume: x-soft }
```

[39] © Robert Tolksdorf, Berlin

Zusammenfassung

- CSS trennt Inhalt und Darstellung
- CSS ist getrennte Sprache für Darstellungseigenschaften
- Vielfältige Einstellungsmöglichkeiten
- CSS2: Erweiterung auf andere Medienarten

[40] © Robert Tolksdorf, Berlin

XML und Darstellung

XML+CSS

- Cascading Style Sheets definieren Darstellung von Tags durch Belegen von CSS-Attributen
- Während ursprünglich für HTML entworfen auch für XML nutzbar
- Darstellung vom eigenen Element `<price>` weiss auf schwarz:

```
price {
    color: white;
    background-color: black;
}
```
- CSS Attribute für visuelle oder auditive Ausgabe von Texten geeignet
- www.w3.org/1999/06/REC-xml-stylesheet-19990629

XSL

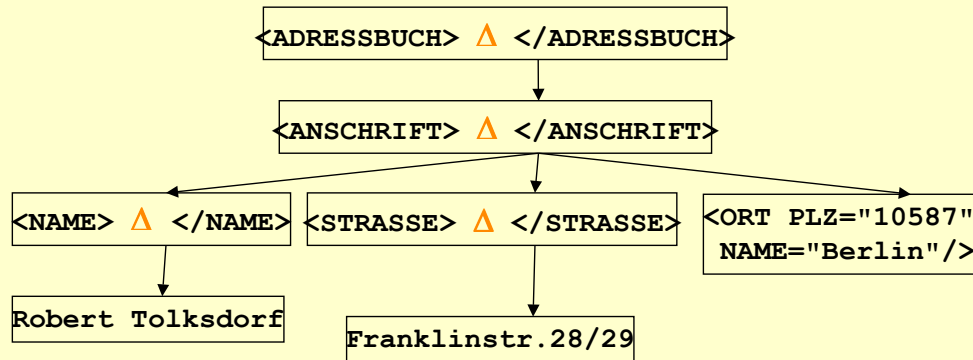
- CSS Vorgehen:
 - HTML enthält Struktur in Inhalt
 - CSS definiert Darstellungseigenschaften
 - Struktur fest
- XSL Standard
 - XML enthält Inhalt
 - XSL transformiert in Darstellungsstruktur (und deren Eigenschaften)

XSL Standards

- Extensible Stylesheet Language (XSL)
 - XSL Transformations (XSLT)
 - Zweck: Transformation von XML Dokumenten
 - Status: W3C Recommendation 16 November 1999
 - Quelle: <http://www.w3.org/TR/xslt>
 - XML Path Language (XPath)
 - Zweck: Ausdrücke mit denen Stellen im XML Dokument bezeichnet werden können
 - Status: W3C Recommendation 16 November 1999
 - Quelle: <http://www.w3.org/TR/xpath>
 - XSL Formatting Objects
 - Zweck: Pseudobäume mit Darstellungseigenschaften
 - Status: W3C Recommendation 15 October 2001
 - Quelle: <http://www.w3.org/TR/xsl/>

XSL(T)

- XML-Dokumente sind Bäume:



- XSLT ist eine Sprache zur Transformation eines XML-Baums in einen anderen (www.w3.org/Style/XSL)

[45] © Robert Tolksdorf, Berlin

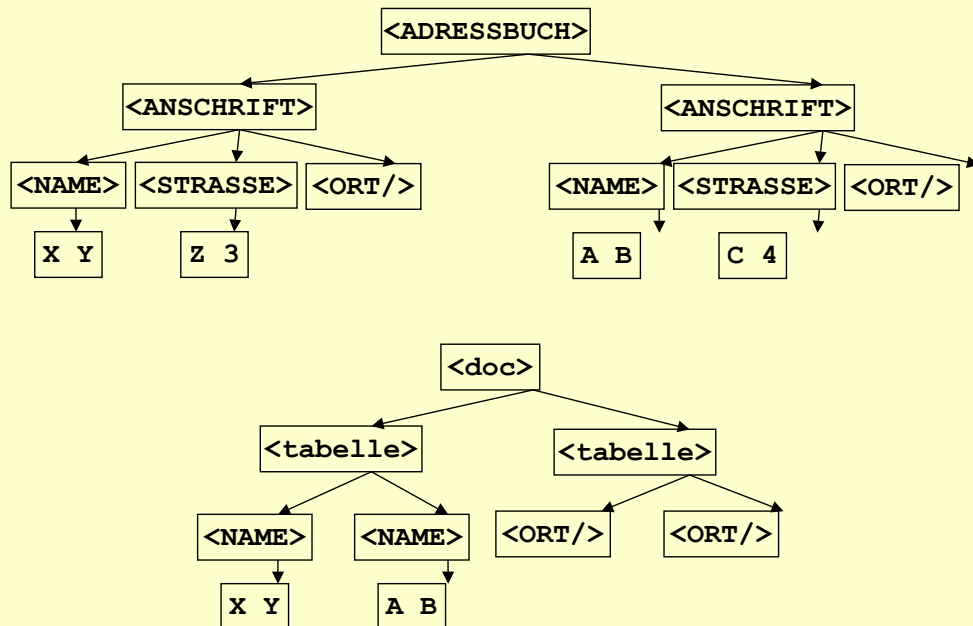
XSL Regeln

- Eine XSL-Regel definiert Muster und dazugehörige Transformationen
- Sie wird auf alle passenden Knoten im Quelldokument angewandt
- Beispiel:

```
<xsl:template match="ANSCHRIFT">
  <tabelle>
    <xsl:apply-templates select="NAME"/>
  </tabelle>
  <tabelle>
    <xsl:apply-templates select="ORT"/>
  </tabelle>
</xsl:template>
```

[46] © Robert Tolksdorf, Berlin

Baumtransformation



[47] © Robert Tolksdorf, Berlin

XSLT

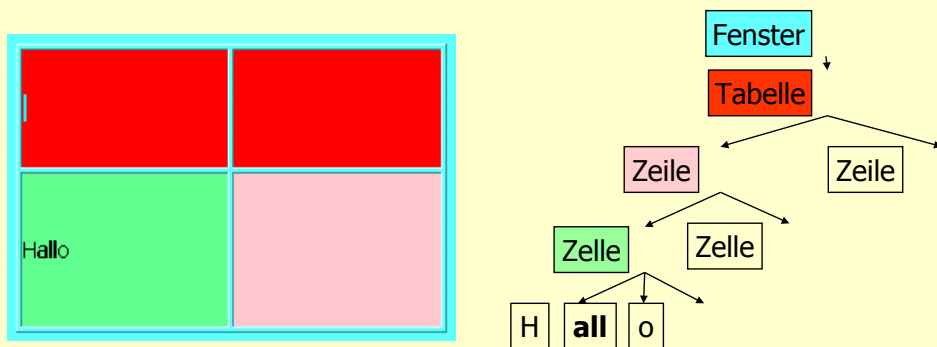
- Zählen aller Kinder eines Knotens:

```
<?xml version="1.0"?>
<xsl:stylesheet version="1.0" xmlns:xsl=
  "http://www.w3.org/1999/XSL/Transform">
  <xsl:template match="/*//*">
    <xsl:value-of
      select="count(descendant::*)" />
    Elemente
  </xsl:template>
</xsl:stylesheet>
```

[48] © Robert Tolksdorf, Berlin

Formatting Objects

- Bildschirmdarstellung ist auch Baum:



- Formatting Objects definiert Knoten und Attribute als Ziel einer Transformation
- Als "Nebeneffekt": Bildschirmdarstellung, PDF Generierung

[49] © Robert Tolksdorf, Berlin

Zusammenfassung

- XSLT transformiert XML Bäume in andere XML Bäume
- Darstellung als „Nebeneffekt“ in Formatting Objects

[50] © Robert Tolksdorf, Berlin

Mehrsprachigkeit im Web Zeichen, Schriften, Sprachen

- Unterschiede in
 - Sprache
 - Schriftzeichen
 - Schriftcodierung
 - Schreibrichtung
 - Kulturellen Konventionen
 - ...



[51] © Robert Tolksdorf, Berlin

[52] © Robert Tolksdorf, Berlin

Sprachen der Welt

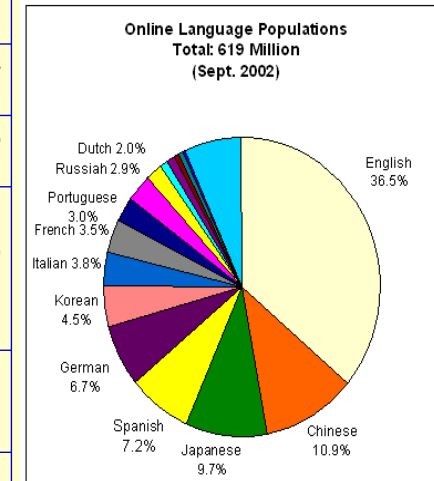
Sprache	Familie	Hauptgebiet	~Mio. Sprechende
Chinesisch	Sinotibetisch	China	885
Englisch	Indoeuropäisch (Germanische Gruppe)	Nordamerika, Großbritannien, Australien, Südafrika	450
Hindi-Urdu	Indoeuropäisch (Indoiranische Gr.)	Indien, Pakistan	333
Spanisch	Indoeuropäisch (Romanische Gr.)	Südamerika, Spanien	266
Portugiesisch	Indoeuropäisch (Romanische Gr.)	Brasilien, Portugal, Angola, Mosambik	175
Bengali	Indoeuropäisch (Indoiranische Gr.)	Bangladesh, Indien	162
Russisch	Indoeuropäisch (Slawische Gr.)	ehem. UdSSR	153
Arabisch	Nordafrikanisch	Afroasiatisch, Naher Osten	150
Japanisch	Altisch	Japan	126
Französisch	Indoeuropäisch (Romanische Gr.)	Frankreich, Kanada, Belgien, Schweiz, Schwarzafrika	122
Deutsch	Indoeuropäisch (Germanische Gr.)	Deutschland, Österreich, Schweiz	118
Wu	Sinotibetisch	China (Schanghai)	77
Javanisch	Austronesisch	Indonesien (Java)	75
Koreanisch	Altisch	Korea	72
Italienisch	Indoeuropäisch (Romanische Gr.)	Italien	63
Marathi	Indoeuropäisch (Indoiranische Gr.)	Südindien	65
Telugu	Drawidisch	Südindien	55
Tamil	Drawidisch	Südindien, Sri Lanka	48
Kantonesisch	Sinotibetisch	China (Kanton)	47
Ukrainisch	Indoeuropäisch (Slawische Gr.)	Ukraine	46

Quelle: Ethnologue, 12. Ausgabe, Dallas, Texas, USA, 1992, nach <http://babel.alis.com:8080/langues/grandes.htm>

[53] © Robert Tolksdorf, Berlin

Sprachen der Online-Population

	Internet access (M)	%'age world onl. pop	Total pop. (M)	%'age of world econ.
English	230.6	36.5%	508	33.4 %
<i>Non-English</i>	403.5	63.5%	5633	66.6 %
TOTAL EUROPEAN LANGUAGES (excl. English)	224.1	35.5%	1218	33.9 %
TOTAL ASIAN LANGUAGES	179.4	28.3%		
TOTAL WORLD	619.0		6200	



[Quelle: Global Reach (global-reach.biz/globstats), 9/02]

[54] © Robert Tolksdorf, Berlin

Internationalisierung

- *Internationalisierung* ist die Planung und Implementierung von Diensten und Produkten so dass sie einfach an lokale Sprachen und Kulturen anpassbar sind, was *Lokalisierung* ist
- Internationalisierung
 - „I18N“ - „I - eighteen letters –N“ – „Internationalization“
 - Voraussetzung für Lokalisierung
 - Beispiele
 - Platzgestaltung in GUIs läßt Raum für Sprachen die mehr Zeichen benötigen
 - Verwendung internationaler Zeichenrepertoires und –codes, z.B. Unicode
 - Vergabe leicht übersetzbarer Beschreibungen für Graphiken
 - Verwendung allgemeinverständlicher Beispiele (Social Security Number ...)
 - Vorausplanung der Übersetzung in Sprachen mit Kodierungen mit mehr als einem Byte pro Zeichen in Software

[55] © Robert Tolksdorf, Berlin

Lokalisierung

- *Lokalisierung* ist die Anpassung eines Produktes oder Dienstes an eine Sprache, Kultur und lokales “look-and-feel” was durch *Internationalisierung* vereinfacht wird
- Lokalisierung
 - „L10N“ – „L - ten letters –N“ – „Localization“
 - Übersetzung
 - Aber auch: Anpassung an Zeitzonen, Währung, Feiertage, Farbkonventionen, Namen, Geschlechterrollen etc.
 - Ziel: Lokalisiertes Produkt oder Dienst soll so aussehen, als sei er/es lokal entwickelt worden

[56] © Robert Tolksdorf, Berlin

Bezeichnung der Sprache

Sprachbezeichner

- Sprachen im Internet durch Codes bezeichnet
- Basis nach RFC 3066 (früher 1766)
 - In ISO 639 definierte Kürzel für Sprachen
 - In ISO 3166 definierte Kürzel für Länder
- Format
 - Sprachcode: de en etc.
 - Sprachcode-Ländercode: de-ch en-uk
 - Matching nach Substring am Anfang en passt auf en-us
 - Groß-/Kleinschreibung irrelevant en passt auf En-us und EN
 - Experimentell: x-kl i ngon
(siehe auch <http://www.google.com/intl/xx-kl i ngon/>)
- Nicht perfekt: Lateinamerikanisches Spanisch?

Sprachcodes nach ISO 639

aa	Afar	eu	Baskisch	kl	Grönländisch	or	Orija	ta	Tamilisch
ab	Abchasisch	fa	Persisch	km	Kambodschanisch	pa	Pundjabisch	te	Telugu
af	Afrikaans	fi	Finnisch	kn	Kannada	pl	Polnisch	tg	Tadschikisch
am	Amharisch	fj	Fiji	ko	Koreanisch	ps	Paschtu	th	Thai
ar	Arabisch	fo	Faröisch	ks	Kaschmirisch	pt	Portugiesisch	ti	Tigrinja
as	Assamesisch	fr	Französisch	ku	Kurdisch	qu	Quechua	tk	Turkmenisch
ay	Aymara	fy	Friesisch	ky	Kirgisisch	rm	Rätoromanisch	tl	Tagalog
az	Aserbaidshianisch	ga	Irish	la	Lateinisch	rn	Kirundisch	tn	Sezuan
ba	Baschkirisch	gd	Schottisches Gälisch	ln	Lingalisch	ro	Rumänisch	to	Tongaisch
be	Belorussisch	gl	Galizisch	lo	Laotisch	ru	Russisch	tr	Türkisch
bg	Bulgarisch	gn	Guarani	lt	Litauisch	rw	Kijarwanda	ts	Tsongaisch
bh	Biharisch	gu	Gujaratisch	lv	Lettisch	sa	Sanskrit	tt	Tatarisch
bi	Bislamisch	ha	Hausa	mg	Malagasisch	sd	Zinti	tw	Twi
bn	Bengalisch	he	(iw) Hebräisch	mi	Maorisch	sg	Sango	uk	Ukrainisch
bo	Tibetisch	hi	Hindi	mk	Mazedonisch	sh	Serbokroatisch	ur	Urdu
br	Bretonisch	hr	Kroatisch	ml	Malajalam	si	Singhalesisch	uz	Usbekisch
ca	Katalanisch	hu	Ungarisch	mn	Mongolisch	sk	Slowakisch	vi	Vietnamesisch
co	Korsisch	hy	Armenisch	mo	Moldavisch	sl	Slowenisch	vo	Volapük
cs	Tschechisch	ia	Interlingua	mr	Marathi	sm	Samoanisch	wo	Wolof
cy	Walisisch	id	(in) Indonesisch	ms	Malaysisch	sn	Schonisch	xh	Xhosa
da	Dänisch	ie	Interlingue	mt	Maltesisch	so	Somalisch	yi	(ji) Jiddish
de	Deutsch	ik	Inupiak	my	Burmesisch	sq	Albanisch	yo	Joruba
dz	Bhutani	is	Isländisch	na	Nauruisch	sr	Serbisch	zh	Chinesisch
e1	Griechisch	it	Italienisch	ne	Nepalisch	ss	Swasiländisch	zu	Zulu
en	Englisch	ja	Japanisch	nl	Holländisch	st	Sesothisch		
eo	Esperanto	jw	Javanisch	no	Norwegisch	su	Sudanesisch		
es	Spanisch	ka	Georgisch	oc	Okzitanisch	sv	Schwedisch		
et	Estnisch	kk	Kasachisch	om	Oromo	sw	Suaheli		

Ländercodes nach ISO 3166

AFGHANISTAN	AF	BRITISH INDIAN OCEAN TERRITORY	IO
ALBANIA	AL	BRUNEI DARUSSALAM	BN
ALGERIA	DZ	BULGARIA	BG
AMERICAN SAMOA	AS	BURKINA FASO	BF
ANDORRA	AD	BURUNDI	BI
ANGOLA	AO	CAMBODIA	KH
ANGUILLA	AI	CAMEROON	CM
ANTARCTICA	AQ	CANADA	CA
ANTIGUA AND BARBUDA	AG	CAPE VERDE	CV
ARGENTINA	AR	CAYMAN ISLANDS	KY
ARMENIA	AM	CENTRAL AFRICAN REPUBLIC	CF
ARUBA	AW	CHAD	TD
AUSTRALIA	AU	CHILE	CL
AUSTRIA	AT	CHINA	CN
AZERBAIJAN	AZ	CHRISTMAS ISLAND	CX
BAHAMAS	BS	COCOS (KEELING) ISLANDS	CC
BAHRAIN	BH	COLOMBIA	CO
BANGLADESH	BD	COMOROS	KM
BARBADOS	BB	CONGO	CG
BELARUS	BY	CONGO, THE DEMOCRATIC REPUBLIC OF THE	CD
BELGIUM	BE	COOK ISLANDS	CK
BELIZE	BZ	COSTA RICA	CR
BENIN	BJ	CÔTE D'IVOIRE	CI
BERMUDA	BM	CROATIA	HR
BHUTAN	BT	CUBA	CU
BOLIVIA	BO	CYPRUS	CY
BOSNIA AND HERZEGOVINA	BA	CZECH REPUBLIC	CZ
BOTSWANA	BW	DENMARK	DK
BOUVET ISLAND	BV	DJIBOUTI	DJ
BRAZIL	BR	DOMINICA	DM

Ländercodes nach ISO 3166

DOMINICAN REPUBLIC	DO	GUINEA-BISSAU	GW
EAST TIMOR	TL	GUYANA	GY
ECUADOR	EC	HAITI	HT
EGYPT	EG	HEARD ISLAND AND MCDONALD ISLANDS	HM
EL SALVADOR	SV	HOLY SEE (VATICAN CITY STATE)	VA
EQUATORIAL GUINEA	GQ	HONDURAS	HN
ERITREA	ER	HONG KONG	HK
ESTONIA	EE	HUNGARY	HU
ETHIOPIA	ET	ICELAND	IS
FALKLAND ISLANDS (MALVINAS)	FK	INDIA	IN
FAROE ISLANDS	FO	INDONESIA	ID
FIJI	FJ	IRAN, ISLAMIC REPUBLIC OF	IR
FINLAND	FI	IRAQ	IQ
FRANCE	FR	IRELAND	IE
FRENCH GUIANA	GF	ISRAEL	IL
FRENCH POLYNESIA	PF	ITALY	IT
FRENCH SOUTHERN TERRITORIES	TF	JAMAICA	JM
GABON	GA	JAPAN	JP
GAMBIA	GM	JORDAN	JO
GEORGIA	GE	KAZAKHSTAN	KZ
GERMANY	DE	KENYA	KE
GHANA	GH	KIRIBATI	KI
GIBRALTAR	GI	KOREA, DEMOCRATIC PEOPLE'S REPUBLIC OF	KP
GREECE	GR	KOREA, REPUBLIC OF	KR
GREENLAND	GL	KUWAIT	KW
GRENADA	GD	KYRGYZSTAN	KG
GUADELOUPE	GP	LAO PEOPLE'S DEMOCRATIC REPUBLIC	LA
GUAM	GU	LATVIA	LV
GUATEMALA	GT	LEBANON	LB
GUINEA	GN	LESOTHO	LS

[61] © Robert Tolksdorf, Berlin

Ländercodes nach ISO 3166

LIBERIA	LR	NETHERLANDS	NL
LIBYAN ARAB JAMAHIRIYA	LY	NETHERLANDS ANTILLES	AN
LIECHTENSTEIN	LI	NEW CALEDONIA	NC
LITHUANIA	LT	NEW ZEALAND	NZ
LUXEMBOURG	LU	NICARAGUA	NI
MACAO	MO	NIGER	NE
MACEDONIA, THE FORMER YUGOSLAV REPUBLIC OF	MK	NIGERIA	NG
MADAGASCAR	MG	NIUE	NU
MALAWI	MW	NORFOLK ISLAND	NF
MALAYSIA	MY	NORTHERN MARIANA ISLANDS	MP
MALDIVES	MV	NORWAY	NO
MALI	ML	OMAN	OM
MALTA	MT	PAKISTAN	PK
MARSHALL ISLANDS	MH	PALAU	PW
MARTINIQUE	MQ	PALESTINIAN TERRITORY, OCCUPIED	PS
MAURITANIA	MR	PANAMA	PA
MAURITIUS	MU	PAPUA NEW GUINEA	PG
MAYOTTE	YT	PARAGUAY	PY
MEXICO	MX	PERU	PE
MICRONESIA, FEDERATED STATES OF	FM	PHILIPPINES	PH
MOLDOVA, REPUBLIC OF	MD	PITCAIRN	PN
MONACO	MC	POLAND	PL
MONGOLIA	MN	PORTUGAL	PT
MONTSERRAT	MS	PUERTO RICO	PR
MOROCCO	MA	QATAR	QA
MOZAMBIQUE	MZ	RÉUNION	RE
MYANMAR	MM	ROMANIA	RO
NAMIBIA	NA	RUSSIAN FEDERATION	RU
NAURU	NR	RWANDA	RW
NEPAL	NP	SAINT HELENA	SH

[62] © Robert Tolksdorf, Berlin

Ländercodes nach ISO 3166

SAINT KITTS AND NEVIS	KN	THAILAND	TH
SAINT LUCIA	LC	TOGO	TG
SAINT PIERRE AND MIQUELON	PM	TOKELAU	TK
SAINT VINCENT AND THE GRENADINES	VC	TONGA	TO
SAMOA	WS	TRINIDAD AND TOBAGO	TT
SAN MARINO	SM	TUNISIA	TN
SÃO TOME AND PRINCIPE	ST	TURKEY	TR
SAUDI ARABIA	SA	TURKMENISTAN	TM
SENEGAL	SN	TURKS AND CAICOS ISLANDS	TC
SEYCHELLES	SC	TUVALU	TV
SIERRA LEONE	SL	UGANDA	UG
SINGAPORE	SG	UKRAINE	UA
SLOVAKIA	SK	UNITED ARAB EMIRATES	AE
SLOVENIA	SI	UNITED KINGDOM	GB
SOLOMON ISLANDS	SB	UNITED STATES	US
SOMALIA	SO	UNITED STATES MINOR OUTLYING ISLANDS	UM
SOUTH AFRICA	ZA	URUGUAY	UY
SOUTH GEORGIA AND THE SOUTH SANDWICH ISLANDS	GS	UZBEKISTAN	UZ
SPAIN	ES	VANUATU	VU
SRI LANKA	LK	VENEZUELA	VE
SUDAN	SD	VIET NAM	VN
SURINAME	SR	VIRGIN ISLANDS, BRITISH	VG
SVALBARD AND JAN MAYEN	SJ	VIRGIN ISLANDS, U.S.	VI
SWAZILAND	SZ	WALLIS AND FUTUNA	WF
SWEDEN	SE	WESTERN SAHARA	EH
SWITZERLAND	CH	YEMEN	YE
SYRIAN ARAB REPUBLIC	SY	YUGOSLAVIA	YU
TAIWAN, PROVINCE OF CHINA	TW	ZAMBIA	ZM
TAJKISTAN	TJ	ZIMBABWE	ZW
TANZANIA, UNITED REPUBLIC OF	TZ		

[63] © Robert Tolksdorf, Berlin

Markierung sprachlicher Eigenschaften


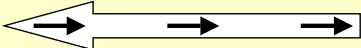
[64] © Robert Tolksdorf, Berlin

Spracheigenschaften in HTML

- Alle HTML Elemente können Sprachbezogene Attribute tragen
 - lang-Attribut: Wert ist Sprachcode
 - Wird vom umgebenden Element „geerbt“
 - Kann jeweils überschrieben werden
 - Default ist durch Content-language HTTP Header gegeben
 - dir-Attribut: (Horizontale) Schreibrichtung der Schrift
 - ltr: Left-to-Right
 - rtl: Right-to-Left
 - Wird vom umgebenden Element „geerbt“
 - Kann jeweils überschrieben werden

[65] © Robert Tolksdorf, Berlin

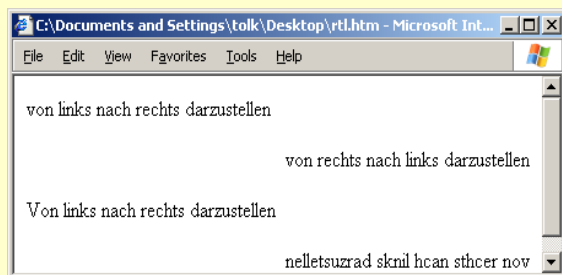
Spracheigenschaften in HTML / Schriftrichtung

- ABC, DEF, GHI aus Schrift, die rechts nach links geschrieben wird (mit `` markiert)
- RST, UVW aus Schrift, die links nach rechts geschrieben wird (mit `` markiert)
- `ABC RST DEF UVW GHI`
- Zwei Möglichkeiten, UNICODE Bidirectional Algorithm
 - `<html dir=„ltr“>`:
ABC TSR DEF **WVU** GHI
 „Embedding“
 - `<html dir=„rtl“>`:
GHI **WVU** DEF TSR ABC
(Schematisch, Details abhängig von Sprachidentifikation, Zeichen etc.)

[66] © Robert Tolksdorf, Berlin

Spracheigenschaften in HTML / Schriftrichtung

- Was ist die Schreibrichtung des Quelltextes?
- Falls schon visuell geordnet, dann versagt Verarbeitung der Richtungsangaben
- Bidirectional Algorithm Override, `<bdo>`-Tag:
 - `<p dir="ltr">von links nach rechts darzustellen</p>`
 - `<p dir="rtl">von rechts nach links darzustellen</p>`
 - `<p dir="ltr"><bdo>von links nach rechts darzustellen</bdo></p>`
 - `<p dir="rtl"><bdo>von rechts nach links darzustellen</bdo></p>`



(Beispiel ist deutschsprachig, daher nur andere Ausrichtung bei `dir="rtl"`)

[67] © Robert Tolksdorf, Berlin

Spracheigenschaften in HTML / Erklärungen

- „Ruby“ ist erklärende Annotation für einen anderen Text
 - World Wide Web ← *ruby text* 新幹線 ← *ruby base* しんかんせん ← *ruby text*
 - W W W ← *ruby base* shinkansen ← *ruby text* 新幹線 ← *ruby base*
- ```
<ruby>
 <rb>www</rb>
 <rt>world wide web</rt>
</ruby>
```

[68] © Robert Tolksdorf, Berlin

## Spracheigenschaften in CSS

- In CSS2 neue Pseudoklasse :lang

```
:lang(en) {color: red}
:lang(fr) {color: blue}
```

  - Noch nicht implementiert
- In CSS2 Selektorausdrücke auf Inhalt des lang Attributs

```
*[lang|=en] {color: red}
*[lang|=fr] {color: blue}
```

Ein Absatz mit einem **chaotic** Sprachgebrauch **ridicule**.  

```
<p>Ein Absatz mit einem chaotic
Sprachgebrauch ridicule.</p>
```
- Eigenschaft direction mit Werten ltr und rtl
- Eigenschaft unicode-bidi
  - Werte normal, embed, bidi-override
  - <bdo>=unicode-bidi: bidi-override

[69] © Robert Tolksdorf, Berlin

## CSS2: Anführungszeichen

- Eigenschaft quotes legt doppelte und einfache An- und Abführungszeichen fest
- Kombiniert mit lang Pseudoelementen:

```
Q:lang(en) { quotes: ' ' ' ' ' ' ' ' }
Q:lang(no) { quotes: "«" "»" "<" ">" }
```
- Als open-quote und close-quote verwendbar:

```
Q:before { content: open-quote }
Q:after { content: close-quote }
```
- Sprachabhängige Zitatmarkierung:

```
<HTML lang="no">
<HEAD>...</HEAD>
<BODY>
 <P><Q>Trøndere gråter når <Q>vinsjan på kaia</Q>
 blir deklamert.</Q>
</BODY>
</HTML>
```

«Trøndere gråter når <Vinsjan på kaia> blir deklamert.»

[70] © Robert Tolksdorf, Berlin

## Sprachabhängige Anführungszeichen

- »Dansk ´da´ Dänisch«
- „Deutsch `de`“
- “English `en` Englisch”
- « Français « fr » Französisch »
- «Italiano «it» Italienisch»
- «Norsk ´no´ Norwegisch»
- «„ru“ Russisch »

[71] © Robert Tolksdorf, Berlin

## Spracheigenschaften in CSS2

- list-style-type Eigenschaft legt die Nummerierung von Listen fest
- In CSS1  
disc, circle, square, decimal, lower-roman, upper-roman, lower-alpha, upper-alpha, none
- In CSS2 zusätzlich
  - hebrew Hebräisch
  - georgian Georgisch (an, ban, gan, ..., he, tan, in, in-an, ...).
  - hiragana a, i, u, e, o, ka, ki, ...
  - katakana A, I, U, E, O, KA, KI, ...
  - hiragana-iroha i, ro, ha, ni, ho, he, to, ...
  - katakana-iroha I, RO, HA, NI, HO, HE, TO, ...
  - weitere

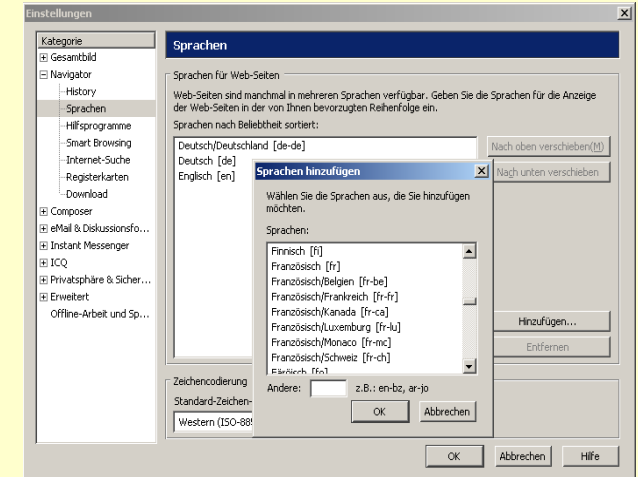
[72] © Robert Tolksdorf, Berlin

## Spracheigenschaften in XML

- Attribut `xml:lang` in XML definiert, also immer verfügbar
- Bedeutung wie `lang` in HTML
- In XHTML sowohl `xml:lang` als auch `lang` benutzen

## Sprache in HTTP

- Browser kann Präferenzen im HTTP-Request mitteilen:  
GET / HTTP/1.1  
AcceptLanguage: en-us;q=0.75,en;q=0.5\*;q=0.25
- `q` gibt Priorität an,  
\* ist Platzhalter
- Vom Browser abhängig:



## Sprache in HTTP

- Server teilt Encoding in Antwort mit  
200 OK HTTP/1.1  
Content-language: fr

<html...

...

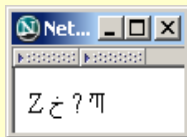
## Zeicheneigenschaften



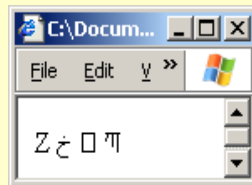
## Zeicheneigenschaften in HTML

- Nicht alle Zeichen müssen darstellbar sein
  - Weil nicht auf lokalen Zeichencode abbildbar
  - Weil keine passende Schrift vorliegt
  - ...
- Vorgehen
  - Nicht vorgeschrieben
  - Teilweise außerhalb der Kontrolle des Darstellers
  - Empfehlung: Visuell klar auf Fehlendes Zeichen hinweisen
  - `&#90;`; `&#xFEA5;` `&#10437;` `&#3904;`

Netscape 7:



IE 5.5:



[81] © Robert Tolksdorf, Berlin

## Zeicheneigenschaften in HTML

- HTML Spezifikation arbeiten immer auf UNICODE Basis
- UNICODE muss nicht benutzt werden, Software muss aber so tun als benutze sie UNICODE
  - Klare Spezifikationen
  - Erlaubt Internationalisierung
  - Unterstützt Lokalisierung
  - Rückwärtskompatibel (ISO 8895-1 gleich unterstem UNICODE Zeichencode)
  - Abstrahiert von Repräsentation der Zeichen in Byteströmen
- UNICODE ist Abstraktionslevel
  - Abstrahiert in der Spezifikation von interner Zeichenkodierung
  - Abstrahiert von Transportrepräsentation

[82] © Robert Tolksdorf, Berlin

## Zeicheneigenschaften

- Zeichenkodierung (Encoding)
  - Character Encoding Form (CEF)
    - Abbildung einer Zeichenfolge auf Strom gleichgroßer Codes
    - z.B. 

005A	FEA5
------	------
  - Character Encoding Scheme (CES)
    - Abbildung einer Zeichenfolge auf einen Bytestrom
    - z.B. 

5A	00	A5	FE
----	----	----	----
- Zeichensatz
  - Bedeutung unklar, kann Repertoire, Code oder Kodierung meinen
- „charset“
  - meint Encoding!

[83] © Robert Tolksdorf, Berlin

## Zeicheneigenschaften in HTML

- Zeichenkodierung
  - Encoding von HTML Dokumenten ist Autoren überlassen
  - Betrifft den Datenstrom zwischen Server und Klienten
  - Klient und Server können die Kodierung aushandeln
  - (Character Set von HTML Dokumenten ist *nicht* verhandelbar)
  - Server und Proxies können Encoding ändern (Transcoding) um Anforderungen des Klienten zu erfüllen
  - Aber: Encoding muss korrekt markiert sein

[84] © Robert Tolksdorf, Berlin

## Zeicheneigenschaften in HTML

- Markierung des Encodings
  - HTTP Header content-Type: text/html; charset=EUC-JP
  - HTML Vorgabe `<meta http-equiv="Content-Type" content="text/html; charset=EUC-JP">`
    - `<meta>` so früh wie möglich im Dokument, bis dahin ASCII
    - charset darf bei Transcoding nicht verändert werden
  - charset Attribut bei HTML Elementen
  - XML Vorgabe `<?xml version="1.0" encoding="UTF-8" ?>`
  - CSS2 Vorgabe `@charset "ISO-8859-1";`

[85] © Robert Tolksdorf, Berlin

## Encoding in HTTP

- Browser kann Präferenzen im HTTP-Request mitteilen:  
GET / HTTP/1.1  
AcceptCharset: iso-8859-1,utf-8;q=0.75,\*;q=0.5
- q gibt Priorität an, \* ist Platzhalter
- Vom Browser abhängig:
  - Microsoft IE: Keine Angabe
  - Netscape 4.72: iso-8859-1,\* ,utf-8
  - NS 6.2: ISO-8859-1, utf-8;q=0.66, \*;q=0.66
  - Opera 6.0:  
windows-1252;q=1.0, utf-8;q=1.0, utf-16;q=1.0, iso-8859-1;q=0.6, \*;q=0.1

[86] © Robert Tolksdorf, Berlin

## Encoding in HTTP

- Server teilt Encoding in Antwort mit  
200 OK HTTP/1.1  
Content-Type: text/html; charset=iso-8859-1  
  
<html...  
...  
▪ ISO-8859-1 als Default bei fehlendem charset vorgesehen
- Praktisch nicht haltbar, weil fehlendes charset andere Ursache haben kann
- Browser muss Encoding bei Darstellung auf System abbilden

[87] © Robert Tolksdorf, Berlin

## Encoding in HTTP

- Bei Formulareingaben entstehen Zeichen, die vom Browser an den Server geschickt werden
- Was ist das Encoding der Eingaben auf dem Weg zum Server?
  - Das Encoding der Seite auf der das Formular stand
  - accept-charset Attribut beim Formular:  
`<form accept-charset="ISO-8859-1, utf-8">`

[88] © Robert Tolksdorf, Berlin

## Transferencoding in HTTP

- Zusätzliche Transferencoding verändert den Inhalt einer übermittelten Information
- Beispiel: Kompimierung durch gzip-Verfahren
- In der Anfrage  
GET / HTTP/1.1  
Accept-Encoding: compress;q=0.5, gzip;q=1.0
- In der Antwort  
200 OK HTTP/1.1  
Content-Encoding: gzip
- Kann auf Transportweg (Proxies) geändert werden

## Content Negotiation

- Auswahl passender Information bezüglich der Dimensionen
  - Medienart (Accept: text/html, text/plain)
  - Sprache (AcceptLanguage: en-us;q=0.75,en;q=0.5;\*;q=0.25)
  - Encoding(Accept-Encoding: compress;q=0.5, gzip;q=1.0)
  - Charset (AcceptCharset: iso-8859-1,utf-8;q=0.75,\*;q=0.5)
  - Angegebene Qualitätsmaße
- Server-abhängige Implementierungen
  - z.B. Schema über Dateinamen:
    - foo.en.html
    - foo.html.en
    - foo.en.html.gz

## Diverse weitere Fragestellungen

- URLs: Momentan nicht Zeichen sondern Byte-basiert
- Wie werden Zeichenketten korrekt gezählt/gemessen?
- Wie werden Zeichenketten normalisiert/verglichen?
- Bezeichner in der Regel nur Teilmenge von ASCII

## Zusammenfassung

- Internationalisierung und Lokalisierung führen zu lokal anpassbaren und angepassten Diensten und Produkten
- Zunehmend bieten
  - HTML/XML
  - CSS
  - HTTP
- Möglichkeiten zur Internationalisierung und Lokalisierung
- Encoding durch charset Parameter und Header
- Sprache durch lang Attribute und Header
- Transferencoding
- Content-Negotiation

## Literatur

- Tex Texin, Yves Savourel. *Tutorial Standards and Practice Web Internationalization*. 2002.  
<http://www.xencraft.com/resources/webi18ntutorial.pdf>
- The Unicode Consortium. *The Unicode Standard, Version 3.0*, Reading, MA, Addison-Wesley Developers Press, 2000.  
<http://www.unicode.org/standard/standard.html>
- H. Alvestrand. Tags for the Identification of Languages. RFC 3066. 2001. <http://www.ietf.org/rfc/rfc3066.txt?number=3066>
- W3C. *Ruby Annotation*. W3C Recommendation 31 May 2001.  
<http://www.w3.org/TR/ruby>
- R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, T. Berners-Lee. RFC 2616 Hypertext Transfer Protocol - HTTP/1.1. June 1999 <ftp://ftp.isi.edu/in-notes/rfc2616.txt>
- Apache HTTP Server Documentation Project. *Apache HTTP Server Version 2.0, Content Negotiation*.  
<http://httpd.apache.org/docs-2.0/content-negotiation.html>

[93] © Robert Tolksdorf, Berlin

## Zusammenfassung

[94] © Robert Tolksdorf, Berlin

## Zusammenfassung

- Darstellung von Inhalten
- Mehrsprachigkeit

[95] © Robert Tolksdorf, Berlin