Netzbasierte Informationssysteme
Mehrsprachigkeit im Web

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

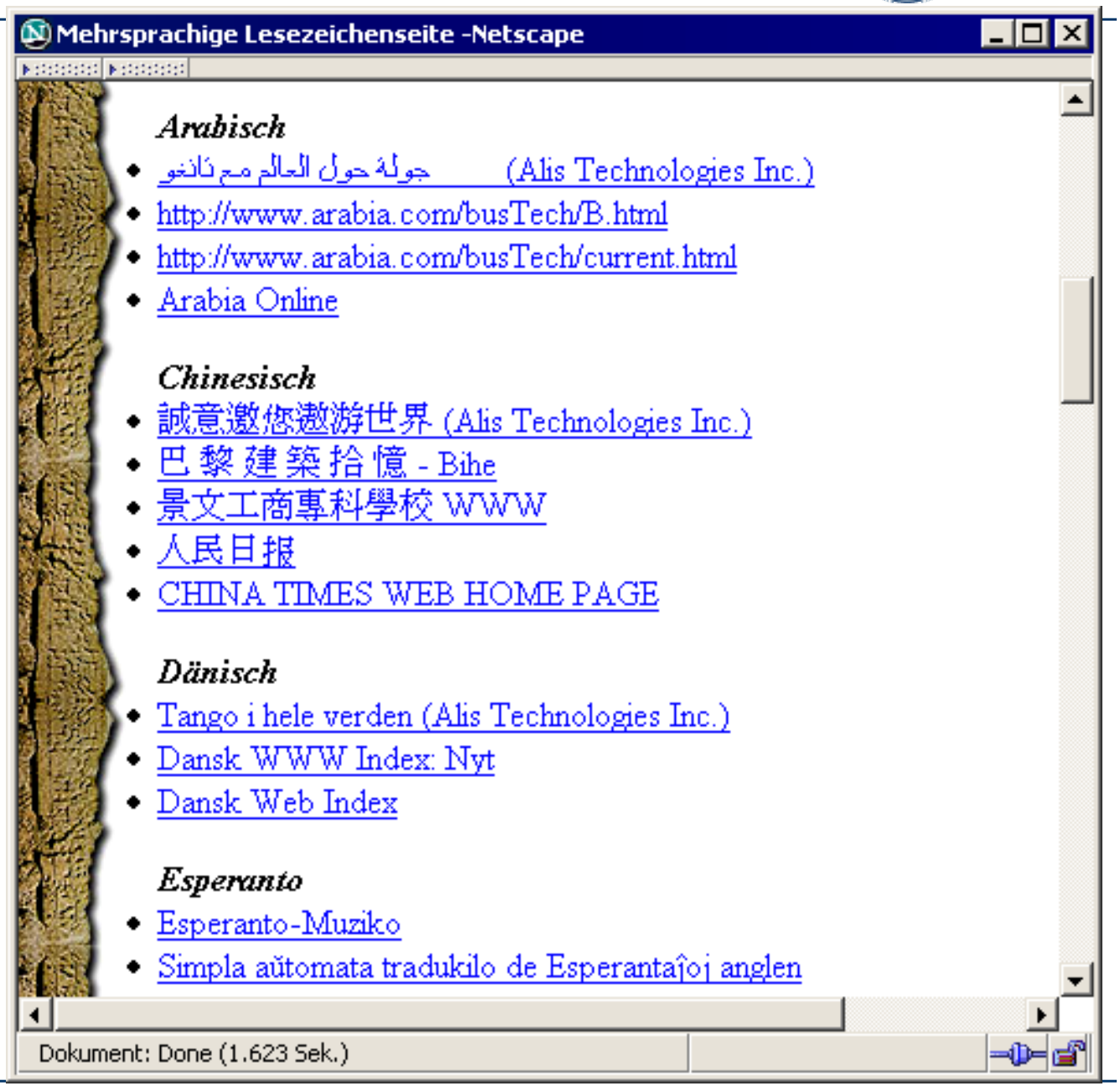
Übersicht

- Sprachen im Web
- Bezeichnung von Sprachen
- Markierung sprachlicher Eigenschaften
- Zeicheneigenschaften

Mehrsprachigkeit im Web
Zeichen, Schriften, Sprachen

Mehrsprachige Seiten

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



Sprachen der Welt

Freie Universität Berlin

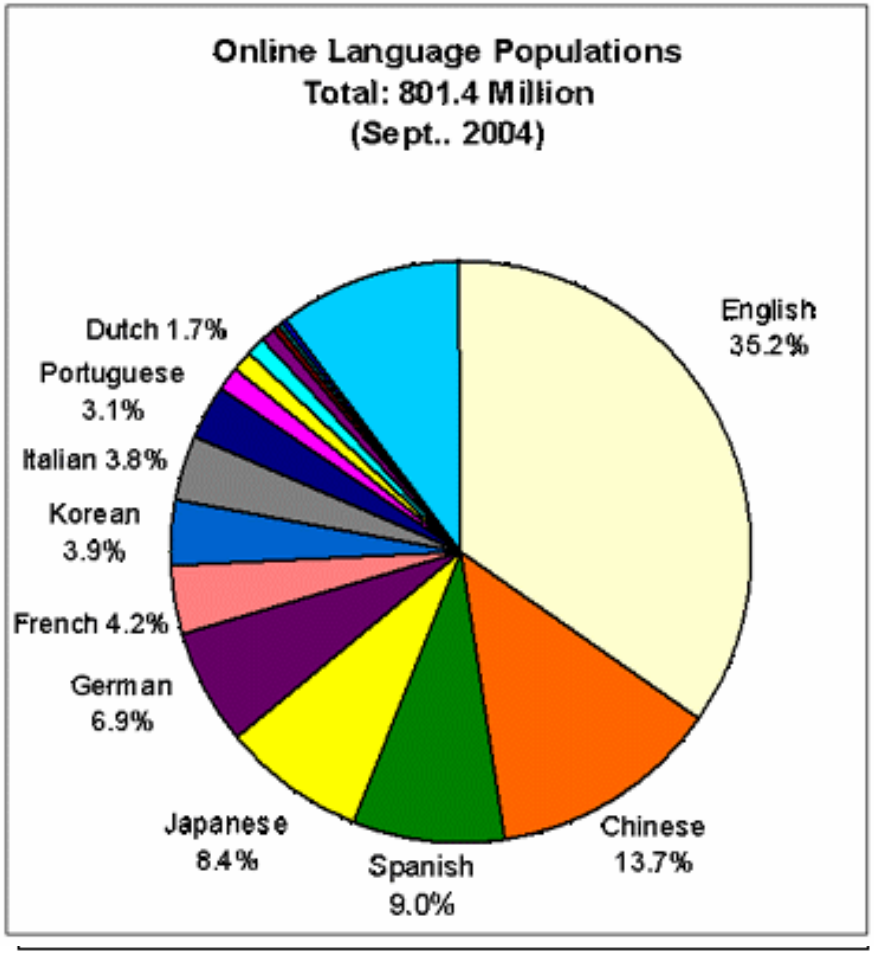
Language	Native	2nd	Total
Mandarin Chinese	873 million	178 million	1051 million
Hindi	370 million	120 million	490 million
English	340 million		510 million
Spanish	350 million	70 million	420 million
Arabic	206 million	24 million	230 million
Portuguese	203 million	10 million	213 million
Bengali	196 million		215 million
Russian	145 million	110 million	255 million
Japanese	126 million	1 million	127 million
German	101 million	128 million	129 million

[Vistawide. http://www.vistawide.com/languages/top_30_languages.htm]

Sprachen der Online-Population

	Internet access (M)	%'age world online population
English	287.5	35.8%
Non-English	516.7	64.2%
TOTAL EUROPEAN LANGUAGES (excl. English)	276.0	37.9%
TOTAL ASIAN LANGUAGES	240.6	33.0%
TOTAL WORLD	729.2	

[Quelldaten nicht konsistent!]



[Quelle: Global Reach (global-reach.biz/globstats), 9/04]

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

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

Bezeichnung von Sprachen

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-klingon
    (siehe auch http://www.google.com/intl/xx-klingon/)
- Nicht perfekt: Lateinamerikanisches Spanisch?

Sprachcodes nach ISO 639

Freie Universität Berlin

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	Aserbaidschanisch	ga	Irisch	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	Haussa	mg	Malagasisch	sd	Zinti	tw	Twi
bn	Bengalisch	he (iw)	Hebräisch	mi	Maorisch	sg	Sango	uk	Ukrainisch
bo	Tibetanisch	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
el	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

Freie Universität Berlin

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

Ländercodes nach ISO 3166

Freie Universität Berlin

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

Ländercodes nach ISO 3166

Freie Universität Berlin

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

Ländercodes nach ISO 3166

Freie Universität Berlin

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

Sprachkürzel nach RFC 4646

- RFC 3066
  - Sprachkürzel:
    Ländercode-Sprachcode (en-US)
  - Bezüge auf ISO Standards
- RFC 4646
  - Sprachkürzel:
    language-script-region-variant-extension-privateuse
  - http://www.iana.org/assignments/language-subtag-registry:
    - Type: language
      Subtag: fr
      Description: French
      Added: 2005-10-16
      Suppress-Script: Latn
    - Type: region
      Subtag: CA
      Description: Canada
      Added: 2005-10-16
  - fr-CA ist gültiges Sprachkürzel

Sprachkürzel nach RFC 4646

- language Kürzel
  - Sprachkürzel, zwei oder drei Buchstaben
- script Kürzel
  - Schreibschrift
  - az-Latn
    Aserbaidschanisch in lateinischer Schrift
- region Kürzel
  - Region
  - es-005: Spanisch in Südamerika
  - zh-Hant-HK: Chinesisch in traditioneller Schreibweise in Hong-Kong
  - Regionen nicht nur Länder (ISO 3166)
  - UNM.49 Codes

Sprachkürzel nach RFC 4646

- Variant Kürzel
  - de-CH-1901: Deutsch nach der Reform von 1901 in der Schweiz…
  - sl-rozaj: Dialekt von Slovenisch
- Extension Kürzel
  - Für spätere Erweiterungen von RFC 4646
- Private-use Kürzel
  - Lokale Kürzel

Markierung sprachlicher Eigenschaften

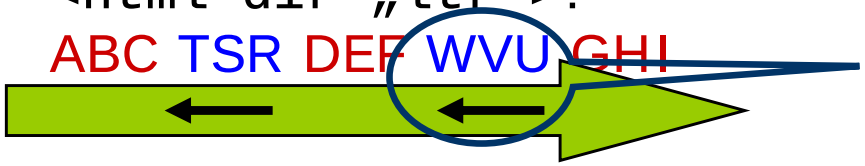
Spracheigenschaften in HTML

- 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

Spracheigenschaften in HTML
Schriftrichtung

Freie Universität Berlin

- ABC, DEF, GHI aus Schrift, die rechts nach links geschrieben wird (mit `<span dir=„ltr">` markiert)
- RST, UVW aus Schrift, die links nach rechts geschrieben wird (mit `<span dir=„rtl">` markiert)
- `<span dir=„ltr">ABC</span> <span dir=„rtl">RST </span> <span dir=„ltr">DEF</span> <span dir=„rtl">UVW</span> GHI`
- Zwei Möglichkeiten, UNICODE Bidirectional Algorithm
  - `<html dir=„ltr">`:
    ABC TSR DEF WVU GHI



„Embedding“

  - `<html dir=„rtl">`:
    GHI WVU DEF TSR ABC

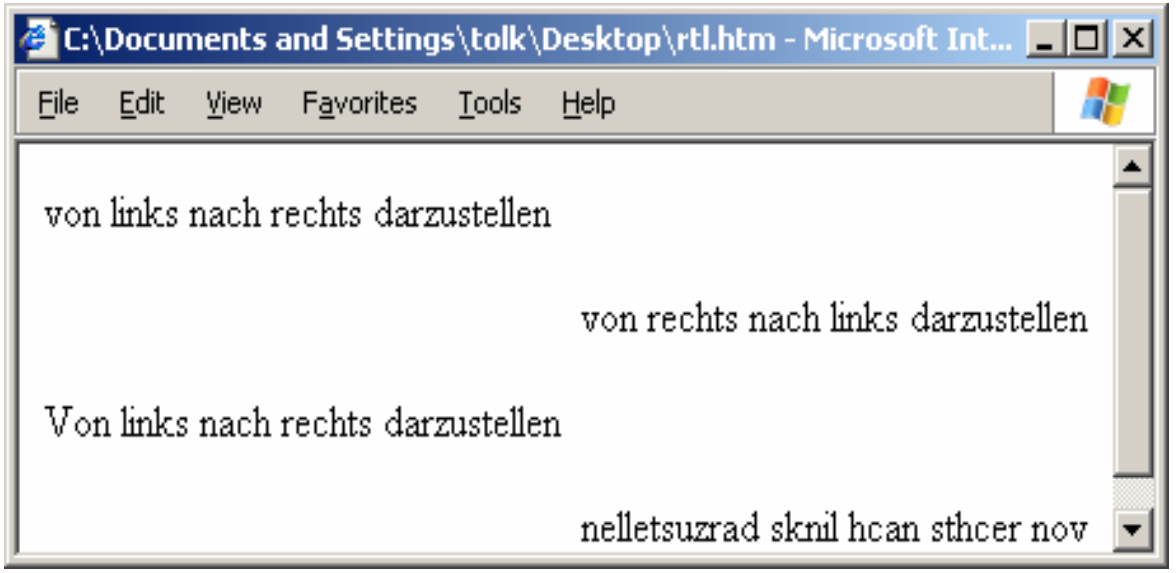
(Schematisch, Details abhängig von Sprachidentifikation, Zeichen etc.)

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>
```

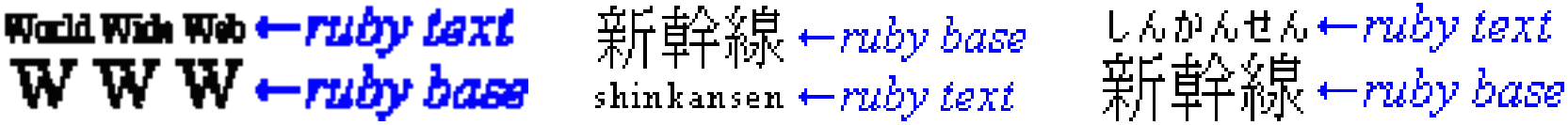
(Beispiel ist deutschsprachig, daher nur andere Ausrichtung bei dir="rtl")



Spracheigenschaften in HTML Erklärungen

Freie Universität Berlin

- „Ruby“ ist erklärende Annotation für einen anderen Text
-




```
<ruby>
  <rb>WWW</rb>
  <rt>World Wide Web</rt>
</ruby>
```

Spracheigenschaften in CSS

- In CSS2 neue Pseudoklasse :lang

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

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

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


```
<p>Ein Absatz mit einem <span lang="en-us">chaotic</span>
Sprachgebrauch <span lang="fr">ridicule</span>.</p>
```

- Eigenschaft `direction` mit Werten `ltr` und `rtl`
- Eigenschaft `unicode-bidi`
  - Werte `normal`, `embed`, `bidi-override`
  - `<bdo>=unicode-bidi: bidi-override`

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.»

Sprachabhängige Anführungszeichen

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

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

Spracheigenschaften in XML

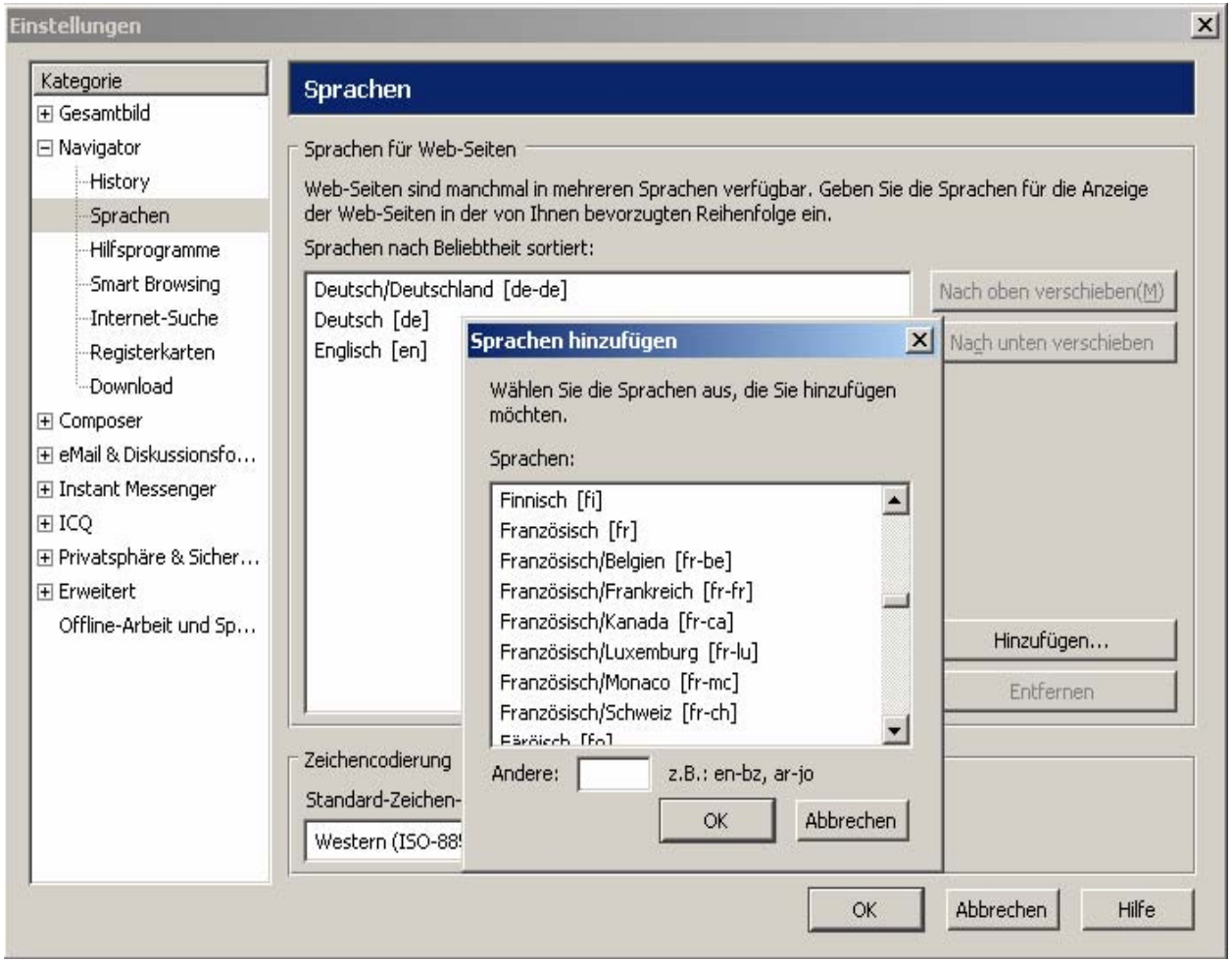
- Attribut `xml:lang` in XML definiert, also immer verfügbar
- Bedeutung wie `lang` in HTML
- In XHTML sowohl `xml:lang` als auch `lang` benutzen

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

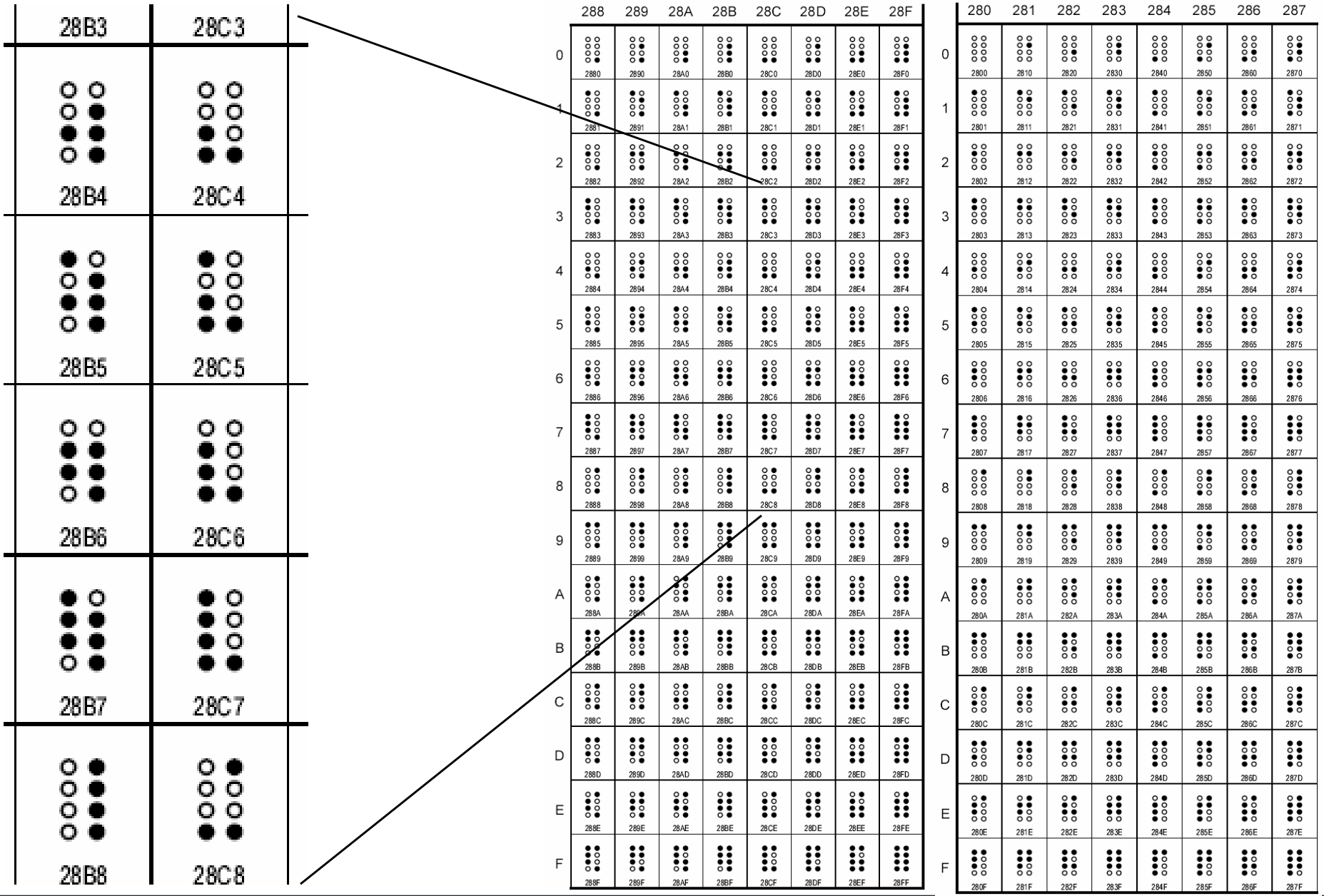
```
200 OK HTTP/1.1
Content-language: fr

<html…
...
```

Zeicheneigenschaften

Zeicheneigenschaften

- Zeichenrepertoire (Character Set, Abstract Character Repertoire, ACS)
  - Eine Menge von Zeichen
  - Definiert durch Namen und Beispiele
  - {Pfund (£), Zett (Z), Ypsilon (Y), Herz (♥)}
  - Keine Ordnung, keine Codierung
- Zeichencode (Coded Character Set, CCS)
  - Abbildung(en) Zeichen → Zeichenposition
  - Z → 5A, ڎ → FEA5 (Khah)
  - z.B. UNICODE, ISO 8859-1

UNICODE: Braille



UNICODE: Beschreibung der Tibetanischen Zeichen

0F36 **Tibetan** **0F7E**

0F36 ⁘ TIBETAN MARK CARET -DZUD RTAGS BZHI MIG CAN
• marks point of text insertion or annotation

0F37 ◌̥ TIBETAN MARK NGAS BZUNG SGOR RTAGS
• emphasis; used like underlining

0F38 ◌̃ TIBETAN MARK CHE MGO

0F39 ◌̱ TIBETAN MARK TSA -PHRU
• a lenition mark

Paired punctuation

0F3A ﴾ TIBETAN MARK GUG RTAGS GYON

0F3B ﴿ TIBETAN MARK GUG RTAGS GYAS
• brackets

0F3C ⟨ TIBETAN MARK ANG KHANG GYON

0F3D ⟩ TIBETAN MARK ANG KHANG GYAS
• used for bracketing with a roof over

Astrological signs

0F3E ◌྾ TIBETAN SIGN YAR TSHES

0F3F ◌྿ TIBETAN SIGN MAR TSHES
• marks which combine with digits

Consonants

0F40 ཀ TIBETAN LETTER KA

0F41 ཁ TIBETAN LETTER KHA

0F42 ག TIBETAN LETTER GA

0F43 གྷ TIBETAN LETTER GHA
≡ 0F42 ག 0FB7 ◌ྷ

0F44 ང TIBETAN LETTER NGA

0F45 ཅ TIBETAN LETTER CA

0F5C ཛྷ TIBETAN LETTER DZHA
≡ 0F5B ཛ 0FB7 ◌ྷ

0F5D ཝ TIBETAN LETTER WA

0F5E ཞ TIBETAN LETTER ZHA

0F5F ཟ TIBETAN LETTER ZA

0F60 འ TIBETAN LETTER -A

0F61 ཡ TIBETAN LETTER YA

0F62 ར TIBETAN LETTER RA
• when followed by a subjoined letter = ra mgo

0F63 ལ TIBETAN LETTER LA

0F64 ཤ TIBETAN LETTER SHA

0F65 ཥ TIBETAN LETTER SSA
= reversed sha

0F66 ས TIBETAN LETTER SA

0F67 ཧ TIBETAN LETTER HA

0F68 ཨ TIBETAN LETTER A
• base for dependent vowels

0F69 ཀྵ TIBETAN LETTER KSSA
≡ 0F40 ཀ 0FB5 ◌ྵ

0F6A ཫ TIBETAN LETTER FIXED-FORM RA
• used only in transliteration and transcription

Dependent vowel signs

0F71 ◌ཱ TIBETAN VOWEL SIGN AA
= a-chung
• common, vowel-lengthening mark

0F72 ◌ི TIBETAN VOWEL SIGN I

0F73 ◌ཱི TIBETAN VOWEL SIGN II
• use of this character is discouraged

Zeicheneigenschaften in HTML

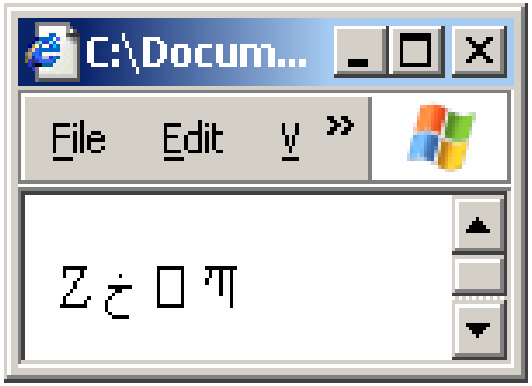
- Das Document Character Set in HTML (ab HTML 4.0)
  - Menge der Zeichen, die in einem HTML Dokument verwendet werden können
  - UNICODE / ISO 10646 („Universal Character Set“, UCS) ist das Document Character Set in HTML, Version 3: >95000 Zeichen
  - Enthält (als Ziel) alle Zeichen der Welt
  - Numerische Zeichenkürzel
    - in HTML `&#90;`
    - HTML und XML: `&#xFEA5;`
    - CSS: `\fea5`
    - Identifizieren ein Zeichen anhand des CCS, in HTML also immer UNICODE
  - HTML Dokumente sind immer in Unicode, egal wie sie später transferiert werden

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:



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

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!

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

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";`

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`

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

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">
```

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

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