



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

MIN-Fakultät
Fachbereich Informatik



Rechnerstrukturen und Betriebssysteme

64-040 Vorlesung RSB
Kapitel 5: Zeichen und Text

Norman Hendrich



Universität Hamburg
Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik
Technische Aspekte Multimodaler Systeme

Wintersemester 2025/2026

Zeichen und Text

- Ad-Hoc Codierungen

- ASCII und ISO-8859

- Unicode

- Tipps und Tricks

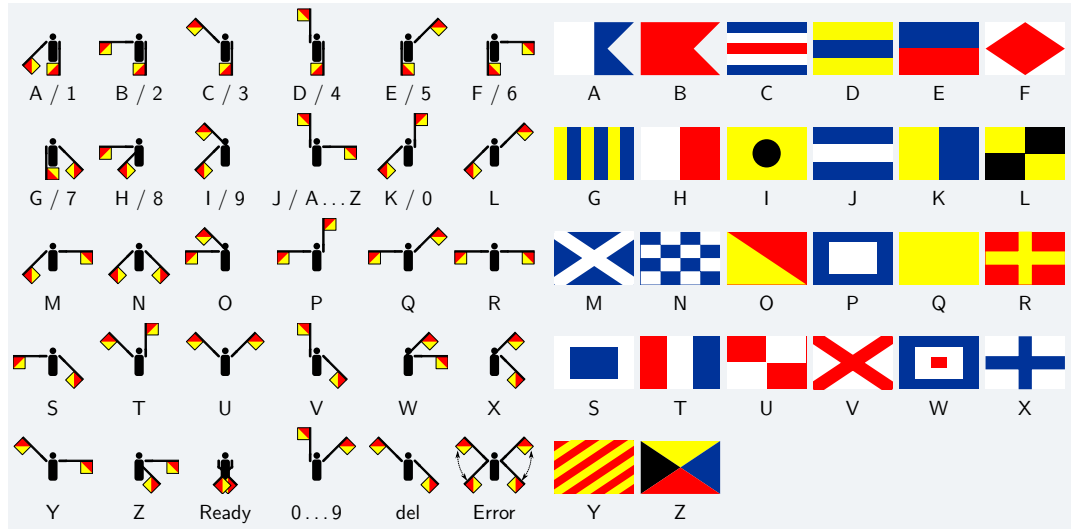
- Base64-Codierung

- Literatur

- ▶ **Zeichen:** engl. *character*
Element z aus einer zur Darstellung von Information vereinbarten, einer Abmachung unterliegenden, endlichen Menge Z von Elementen
- ▶ Die Menge Z heißt **Zeichensatz** oder **Zeichenvorrat** engl. *character set*
- ▶ **Binärzeichen:** engl. *binary element, binary digit, bit*
Jedes der Zeichen aus einem Vorrat / aus einer Menge von zwei Symbolen
- ▶ **Numerischer Zeichensatz:** Zeichenvorrat aus Ziffern und/oder Sonderzeichen zur Darstellung von Zahlen
- ▶ **Alphanumerischer Zeichensatz:** Zeichensatz aus (mindestens) den Dezimalziffern und den Buchstaben des Alphabets, meistens auch mit Sonderzeichen (Leerzeichen, Punkt, Komma usw.)

- ▶ **Alphabet:** engl. *alphabet*
Ein in vereinbarter Reihenfolge geordneter Zeichenvorrat
- ▶ **Zeichenkette:** Eine Folge von Zeichen engl. *string*
- ▶ **Wort:** engl. *word*
Zeichenkette, die in einem gegebenen Zusammenhang als Einheit bezeichnet wird
- ▶ Worte aus 8 Binärzeichen (8 bit) werden als **Byte** bezeichnet
- ▶ **Stelle:** Die Position eines Zeichens in einer Zeichenkette engl. *position*

Flaggen-Signale



- ▶ Symbole als 2x3 Matrix (geprägte Punkte)
- ▶ Erweiterung 2x4 Matrix (für Computer)
- ▶ bis zu 64 (256) mögliche Symbole
- ▶ diverse Varianten
 - ▶ ein Symbol pro Buchstabe
 - ▶ ein Symbol pro Silbe
 - ▶ Kurzschrift/Steno

Gruppe 1

a	b	c	d	e	f	g	h	i	j
●○	●○	●●	●●	●○	●●	●●	●○	○●	○●
○○	●○	○○	○○	○○	○○	○○	○○	○○	○○
○○	○○	○○	○○	○○	○○	○○	○○	○○	○○

Gruppe 2

k	l	m	n	o	p	q	r	s	t
●○	●○	●●	●●	●○	●○	●○	●○	●○	●○
○○	●○	○○	○○	○○	○○	○○	○○	○○	○○
●○	●○	●○	●○	●○	●○	●○	●○	●○	●○

Gruppe 3

u	v	x	y	z			ß	st
●○	●○	●●	●●	●○			○●	○●
○○	●○	○○	○○	○○			●○	●○
●●	●●	●●	●●	●●			●●	●●

Gruppe 4

au	eu	ei	ch	sch		ü	ö	w
●○	●○	●●	●●	●○		●○	○●	○●
○○	●○	○○	○○	○○		●●	●○	●○
○●	○●	○●	○●	○●		○●	○●	○●

Gruppe 5

äu	ä	ie	Zahlz.	Großb.		.	-	'
○●	○●	○●	○●	○●		○○	○○	○○
○○	○●	○○	○○	○○		○○	○○	○○
●○	○○	●●	●●	○●		●○	●●	○●

Gruppe 6

.	:		?	!	()	?	*	^
○○	○○	○○	○○	○○	○○	○○	○○	○○
●○	●○	●○	●○	●○	●○	●○	●○	●○
○○	●○	○○	○○	○○	○○	○○	○○	○○

Codetabelle		• kurzer Ton	– langer Ton
A • –	S • • •	. • – • – • –	S-Start – • – • –
B – • • •	T –	, – – • • – –	Verst. • • • – •
C – • – •	U • • –	? • • – – • •	S-Ende • – • – •
D – • •	V • • • –	' • – – – – •	V-Ende • • • – • –
E •	W • – –	! – • – • – –	Error • • • • • • • •
F • • – •	X – • • –	/ – • • – •	
G – – •	Y – • – –	(– • – – •	Ä • – • –
H • • • •	Z – – • •	) – • – – • –	À • – – • –
I • •	0 – – – – –	& • – • • •	É • • – • •
J • – – –	1 • – – – –	: – – – • • •	È • – • • –
K – • –	2 • • – – –	; – • – • – •	Ö – – – •
L • – • •	3 • • • – –	= – • • • –	Ü • • – –
M – –	4 • • • • –	+ • – • – •	ß • • • – – • •
N – •	5 • • • • •	- – • • • • –	CH – – – –
O – – –	6 – • • • •	_ • • – – • –	Ñ – – • – –
P • – – •	7 – – • • •	" • – • • – •	...
Q – – • –	8 – – – • •	\$ • • • – • • –	
R • – •	9 – – – – •	@ • – • • – •	SOS • • • – – – • • •

ASCII

American Standard Code for Information Interchange

- ▶ eingeführt 1967, aktualisiert 1986: ANSI X3.4-1986
- ▶ viele Jahre der dominierende Code für Textdateien
- ▶ alle Zeichen einer typischen Schreibmaschine
- ▶ Erweiterung des früheren 5-bit Fernschreiber-Codes (Murray-Code)
- ▶ 7-bit pro Zeichen, 128 Zeichen insgesamt
- ▶ 33 Steuerzeichen (engl: *control characters*) 0 ... 0x1F, 0x7F
- ▶ 96 druckbare Zeichen: Buchstaben, Ziffern, Sonderzeichen 0x20 ... 0x7E
 - Leerzeichen 0x20
 - Dezimalziffern 0x30 ... 0x39
 - Großbuchstaben 0x41 ... 0x5A

ASCII: Codetabelle

Code	...0	...1	...2	...3	...4	...5	...6	...7	...8	...9	...A	...B	...C	...D	...E	...F
0...	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1...	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2...	SP	!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
3...	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4...	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5...	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6...	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7...	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL

de.wikipedia.org/wiki/ASCII

- ▶ SP = Leerzeichen, CR = carriage-return, LF = line-feed
- ▶ ESC = escape, DEL = delete, BEL = bell usw.

- ▶ Erweiterung von ASCII um Sonderzeichen und Umlaute
- ▶ 8-bit Codierung: bis max. 256 Zeichen darstellbar
- ▶ Latin-1: Westeuropäisch
- ▶ Latin-2: Mitteleuropäisch
- ▶ Latin-3: Südeuropäisch
- ▶ Latin-4: Baltisch
- ▶ Latin-5: Kyrillisch
- ▶ Latin-6: Arabisch
- ▶ Latin-7: Griechisch
- ▶ usw.
- ▶ immer noch nicht für mehrsprachige Dokumente geeignet

ISO-8859: Sonderzeichen

Sonderzeichen gemeinsam für alle 8859 Varianten

Code	...0	...1	...2	...3	...4	...5	...6	...7	...8	...9	...A	...B	...C	...D	...E	...F
0...	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1...	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2...	wie ISO/IEC 8859, Windows-125X und US-ASCII															
3...																
4...																
5...																
6...																
7...																DEL
8...	PAD	HOP	BPH	NBH	IND	NEL	SSA	ESA	HTS	HTJ	VTS	PLD	PLU	RI	SS2	SS3
9...	DCS	PU1	PU2	STS	CCH	MW	SPA	EPA	SOS	SGCI	SCI	CSI	ST	OSC	PM	APC
A...	wie ISO/IEC 8859-1 und Windows-1252															
B...																
C...																
D...																
E...																
F...																

ISO-8859-1: Codetabelle

Erweiterung von ASCII für westeuropäische Sprachen

1.2 Zeichen und Text - ASCII und ISO-8859

64-040 Rechnerstrukturen und Betriebssysteme

Code	...0	...1	...2	...3	...4	...5	...6	...7	...8	...9	...A	...B	...C	...D	...E	...F
0...	<i>nicht belegt</i>															
1...																
2...	SP	!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
3...	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4...	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5...	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6...	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7...	p	q	r	s	t	u	v	w	x	y	z	{		}	~	
8...	<i>nicht belegt</i>															
9...																
A...	<i>NBSP</i>	ı	ø	£	¤	¥	¦	§	¨	©	ª	«	¬	<i>SHY</i>	®	¯
B...	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
C...	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D...	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
E...	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F...	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

ISO-8859-2

Erweiterung von ASCII für slawische Sprachen

1.2 Zeichen und Text - ASCII und ISO-8859

64-040 Rechnerstrukturen und Betriebssysteme

Code	...0	...1	...2	...3	...4	...5	...6	...7	...8	...9	...A	...B	...C	...D	...E	...F
0...	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1...	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2...	SP	!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
3...	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4...	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5...	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6...	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7...	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL
8...	PAD	HOP	BPH	NBH	IND	NEL	SSA	ESA	HTS	HTJ	VTS	PLD	PLU	RI	SS2	SS3
9...	DCS	PU1	PU2	STS	CCH	MW	SPA	EPA	SOS	SGCI	SCI	CSI	ST	OSC	PM	APC
A...	NBSP	Ą	ˆ	Ł	˝	Ł	Ś	§	˝	Š	Ş	Ť	Ž	SHY	Ž	Ž
B...	°	ą	˘	ł	˛	ł	ś	˜	˘	š	ş	ť	ž	˝	ž	ž
C...	Ř	Á	Â	Ă	Ä	Í	Ć	Ç	Č	É	Ę	Ě	Ě	Í	Î	Ď
D...	Đ	Ń	Ň	Ó	Ô	Õ	Ö	×	Ř	Ů	Ú	Ů	Ü	Ý	Ť	ß
E...	í	á	â	ă	ä	í	ć	ç	č	é	ę	ě	ě	í	î	ď
F...	đ	ń	ň	ó	ô	õ	ö	÷	ř	ů	ú	ů	ü	ý	ț	·

ISO-8859-15

Modifizierte ISO-8859-1 mit € (0xA4)

1.2 Zeichen und Text - ASCII und ISO-8859

64-040 Rechnerstrukturen und Betriebssysteme

Code	...0	...1	...2	...3	...4	...5	...6	...7	...8	...9	...A	...B	...C	...D	...E	...F
0...	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1...	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2...	SP	!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
3...	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4...	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5...	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6...	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7...	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL
8...	PAD	HOP	BPH	NBH	IND	NEL	SSA	ESA	HTS	HTJ	VTS	PLD	PLU	RI	SS2	SS3
9...	DCS	PU1	PU2	STS	CCH	MW	SPA	EPA	SOS	SGCI	SCI	CSI	ST	OSC	PM	APC
A...	NBSP	ı	ç	£	€	¥	Š	§	š	©	ª	«	¬	SHY	®	™
B...	°	±	²	³	Ž	µ	¶	·	ž	¹	º	»	Œ	œ	ÿ	ı
C...	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D...	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
E...	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F...	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

- ▶ Zeichensatz des IBM-PC ab 1981
 - ▶ Erweiterung von ASCII auf einen 8-bit Code
 - ▶ einige Umlaute (westeuropäisch)
 - ▶ Grafiksymbole
-
- ▶ de.wikipedia.org/wiki/Codepage_437
 - ▶ verbesserte Version: Codepage 850, 858 (€-Symbol an 0xD5)
 - ▶ Codepage 1252 entspricht (weitgehend) ISO-8859-1
 - ▶ Sonderzeichen liegen an anderen Positionen als bei ISO-8859

Microsoft: Codepage 850

Code	...0	...1	...2	...3	...4	...5	...6	...7	...8	...9	...A	...B	...C	...D	...E	...F
0...		☺	☹	♥	♦	♣	♠	•	◼	○	◼	♂	♀	♪	♫	☼
1...	▶	◀	↕	!!	¶	§	—	↕	↑	↓	→	←	└	↔	▲	▼
2...		!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
3...	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4...	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5...	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6...	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7...	p	q	r	s	t	u	v	w	x	y	z	{		}	~	△
8...	Ç	ü	é	â	ä	à	â	ç	ê	ë	è	ï	î	ì	Ä	Å
9...	É	æ	Æ	ô	ö	ò	û	ù	ÿ	Ö	Ü	ø	£	Ø	×	f
A...	á	í	ó	ú	ñ	Ñ	ª	º	¿	®	¬	½	¼	¡	«	»
B...	☐	☐	☐		└	Á	Â	À	©	¶		¶	¶	¢	¥	⌞
C...	L	⊥	T	└	—	†	ã	Ã	ℒ	℞	⊥	⊥	⊥	=	⊥	α
D...	ð	Ð	Ê	Ë	È	Ì	Í	Î	Ï	Ɔ	Ɔ	■	■	¡	ì	■
E...	Ó	ß	Ô	Ò	õ	Õ	µ	þ	Ɔ	Ú	Û	Ù	ý	Ý	—	‘
F...		±	=	¾	¶	§	÷	,	°	“	•	¹	³	²	■	

Austausch von Texten?

- ▶ die meisten gängigen Codes (abwärts-) kompatibel mit ASCII
- ▶ unterschiedliche Codierung für Umlaute (soweit vorhanden)
- ▶ unterschiedliche Codierung der Sonderzeichen
- ▶ Systemspezifische Konventionen für Zeilenende
 - ▶ abhängig von Rechner- und Betriebssystem
 - ▶ Konverter-Tools: `dos2unix`, `unix2dos`, `iconv`

Betriebssystem	Zeichensatz	Abkürzung	Hex-Code	Escape
Unix, Linux, Mac OS X, AmigaOS, BSD	ASCII	<i>LF</i>	0A	<code>\n</code>
Windows, DOS, OS/2, CP/M, TOS (Atari)	ASCII	<i>CR LF</i>	0D 0A	<code>\r\n</code>
Mac OS bis Version 9, Apple II	ASCII	<i>CR</i>	0D	<code>\r</code>
AIX OS, OS 390	EBCDIC	<i>NEL</i>	15	

- ▶ zunehmende Vernetzung und Globalisierung
 - ▶ internationaler Datenaustausch?
 - ▶ Erstellung mehrsprachiger Dokumente?
 - ▶ Unterstützung orientalischer oder asiatischer Sprachen?
-
- ▶ ASCII oder ISO-8859-1 reicht nicht aus
 - ▶ temporäre Lösungen konnten sich nicht durchsetzen, z.B: **ISO-2022**
Spezialbefehle zur Umschaltung zwischen mehreren Zeichensätzen,
sog. *Escapesequenzen*
-
- ⇒ **Unicode** als System zur Codierung aller Zeichen aller bekannten Schriftsysteme
auch für tote Schriften/Sprachen

- ▶ auch abgekürzt als UCS: **Universal Character Set**
- ▶ zunehmende Verbreitung (Betriebssysteme, Applikationen)
- ▶ Darstellung erfordert auch entsprechende Schriftarten
- ▶ home.unicode.org www.unicode.org/charts

- ▶ 1991 1.0.0: europäisch, nahöstlich, indisch
- ▶ 1992 1.0.1: ostasiatisch (Han)
- ▶ 1993 akzeptiert als ISO/IEC-10646 Standard
- ...
- ▶ 2024 16.0.0: inzwischen 154 998 Zeichen
 - ▶ Sprachzeichen, Hieroglyphen etc.
 - ▶ Symbole: Satzzeichen, Währungen (\$... ₪), Pfeile, mathematisch, technisch, Braille, Noten etc.
 - ▶ Emojis (3 790 aktuell) / Kombinationen

- ▶ ursprüngliche Version nutzt 16-bit pro Zeichen
 - ▶ die sogenannte „*Basic Multilingual Plane*“
 - ▶ Schreibweise hexadezimal als U+xxxx
 - ▶ Bereich von U+0000 . . . U+FFFF
 - ▶ Schreibweise in Java-Strings: \uxxxx
z.B. \u03A9 für Ω, \u20AC für das €-Symbol
-
- ▶ mittlerweile mehr als 2^{16} Zeichen
 - ▶ Erweiterung um „*Extended Planes*“
 - ▶ U+10000 . . . U+10FFFF

Latin-Zeichen

- ▶ U+0000 bis U+007F: ASCII www.unicode.org/charts/PDF/U0000.pdf
- ▶ U+0100 bis U+017F: Latin-A www.unicode.org/charts/PDF/U0100.pdf
Europäische Umlaute und Sonderzeichen
- ▶ ab U+0180 weitere Sonderzeichen: Latin-B, Latin-C usw.

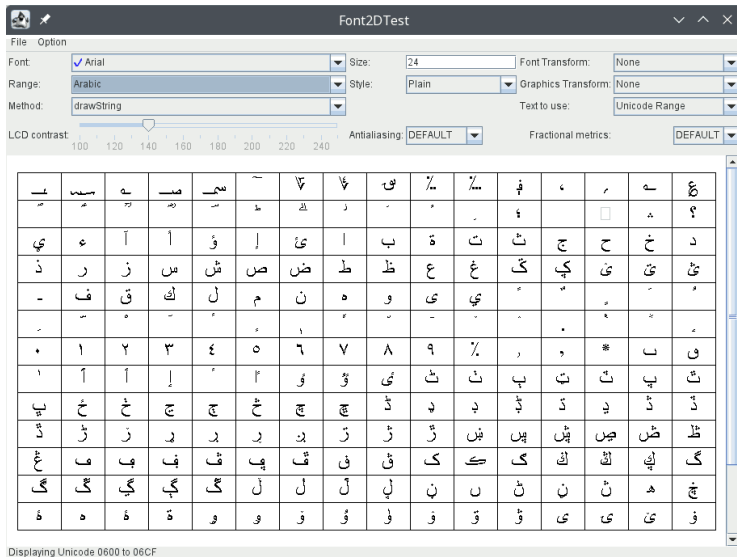
Symbole und Operatoren

- ▶ griechisch www.unicode.org/charts/PDF/U0370.pdf
- ▶ letterlike Symbols www.unicode.org/charts/PDF/U2100.pdf
- ▶ Pfeile www.unicode.org/charts/PDF/U2190.pdf
- ▶ Operatoren www.unicode.org/charts/PDF/U2A00.pdf
- ▶ Dingbats www.unicode.org/charts/PDF/U2700.pdf

Asiatische Sprachen: Chinesisch (traditional/simplified), Japanisch, Koreanisch

- ▶ U+3400 bis U+4DBF www.unicode.org/charts/PDF/U3400.pdf
- ▶ U+4E00 bis U+9FFF www.unicode.org/charts/PDF/U4E00.pdf

Unicode: Java2D Fontviewer



OpenJDK demos and samples

Unicode: in Webseiten (HTML)

- ▶ HTML-Header informiert über verwendeten Zeichensatz
- ▶ Unterstützung und Darstellung abhängig vom Browser
- ▶ Demo kermitproject.org/utf8.html

```
<html>
<head>
<META http-equiv="Content-Type" content="text/html;
      charset=utf-8">
<title>UTF-8 Sampler</title>

<META ...
</head>
...
```

kermitproject.org/utf8.html

1.3 Zeichen und Text - Unicode

64-040 Rechnerstrukturen und Betriebssysteme

I Can Eat Glass

And from the sublime to the ridiculous, here is a **certain phrase**¹ in an assortment of languages:

- [illegible]

- [illegible]

Unicode: Repräsentation?

- ▶ 16-bit für jedes Zeichen, bis zu 65 536 Zeichen
 - ▶ schneller Zugriff auf einzelne Zeichen über Arrayzugriffe (Index)
 - ▶ aber: doppelter Speicherbedarf gegenüber ASCII/ISO-8859-1
 - ▶ Verwendung u.a. in Java: Datentyp `char`
-
- ▶ ab Unicode 3.0 mehrere *Planes* zu je 65 536 Zeichen
 - ▶ direkte Repräsentation aller Zeichen erfordert 32-bit/Zeichen
 - ▶ vierfacher Speicherbedarf gegenüber ASCII/ISO-8859-1
-
- ▶ bei Dateien ist möglichst kleine Dateigröße wichtig
- ⇒ Codierung als UTF-8 oder UTF-16

UTF-8: Codierung

Unicode-Bereich (hexadezimal)	UTF-Codierung (binär)	Anzahl (benutzt)
0000 0000 - 0000 007F	0*** **	128
0000 0080 - 0000 07FF	110* **** 10** ****	1 920
0000 0800 - 0000 FFFF	1110 **** 10** **** 10** ****	63 488
0001 0000 - 0010 FFFF	1111 0*** 10** **** 10** **** 10** ****	bis 2^{21}

- ▶ untere 128 Zeichen kompatibel mit ASCII
- ▶ Sonderzeichen westlicher Sprachen je zwei Bytes
- ▶ führende Eins markiert Multi-Byte Zeichen
- ▶ Anzahl der führenden Einsen gibt Anz. Bytegruppen an
- ▶ Zeichen ergibt sich als Bitstring aus den ***...*
- ▶ theoretisch bis zu sieben Folgebytes a 6-bit: max. 2^{42} Zeichen

UTF-8: Beispiele

Zeichen	Unicode	Unicode binär	UTF-8 binär	UTF-8 hexadezimal
Buchstabe y	U+0079	00000000 01111001	01111001	79
Buchstabe ä	U+00E4	00000000 11100100	11000011 10100100	C3 A4
Zeichen für eingetragene Marke ®	U+00AE	00000000 10101110	11000010 10101110	C2 AE
Eurozeichen €	U+20AC	00100000 10101100	11100010 10000010 10101100	E2 82 AC
Violinschlüssel 	U+1D11E	00000001 11010001 00011110	11110000 10011101 10000100 10011110	F0 9D 84 9E

de.wikipedia.org/wiki/UTF-8

- ▶ Zeichen werden mit variabler Länge codiert, 1 . . . 4-Bytes
- ▶ volle Kompatibilität mit 7-bit ASCII
- ▶ effiziente Codierung von „westlichen“ Unicode-Texten
- ▶ Unterstützung aller bisher definierten Zeichen

Locale: die Sprach-Einstellungen und Parameter

- ▶ auch: `i18n` („internationalization“)
 - ▶ Sprache der Benutzeroberfläche
 - ▶ Tastaturlayout/-belegung
 - ▶ Zahlen-, Währungs-, Datums-, Zeitformate
-
- ▶ Linux/POSIX: Einstellung über die Locale-Funktionen der Standard C-Library (Befehl: `locale`)
 - Java: `java.util.Locale`
 - Windows: Einstellung über System/Registry-Schlüssel

- ▶ `dos2unix`, `unix2dos`: Umwandeln von Dateien (z.B. Programm-Quelltexte) zwischen DOS/Windows und Unix/Linux: Codierung und Zeilenenden

<code>dos2unix -h</code>	Optionen anzeigen / Hilfe
<code>dos2unix -ascii -n a.txt b.txt</code>	nur Umbrüche (von a.txt nach b.txt)
<code>dos2unix -iso -n a.txt b.txt</code>	Umbrüche und ISO-8851-1
<code>unix2dos -850 -n a.txt b.txt</code>	Umbrüche und Codepage 850

- ▶ `iconv`: „Universalwerkzeug“ zur Umwandlung von Textcodierungen

<code>iconv -l</code>	Liste der unterstützten Codierungen
<code>iconv -f <i><encoding></i> ...</code>	Codierung der Eingabedatei
<code>iconv -t <i><encoding></i> ...</code>	Codierung der Ausgabedatei
<code>iconv -o <i><filename></i> ...</code>	Name der Ausgabedatei

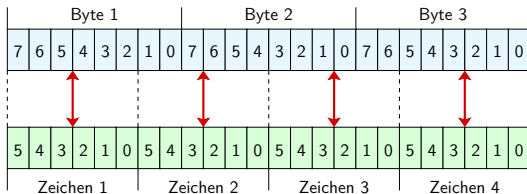
```
iconv -f iso-8859-1 -t utf-8 -o foo.utf8.txt foo.txt
```

- ▶ Konvertierungsfunktionen in den meisten Texteditoren enthalten!

Übertragung von (Binär-) Dateien zwischen verschiedenen Rechnern?

- ▶ SMTP (Internet Mail-Protokoll) verwendet 7-bit ASCII
 - ▶ bei Netzwerk-Übertragung müssen alle Rechner/Router den verwendeten Zeichensatz unterstützen
- ⇒ Verfahren zur Umcodierung der Datei in 7-bit ASCII notwendig
- ⇒ etabliert ist das **Base64** Verfahren (RFC 2045)
- ▶ alle E-Mail Dateianhänge und 8-bit Textdateien
 - ▶ Umcodierung benutzt nur Buchstaben, Ziffern und drei Sonderzeichen
 - ▶ Daten werden byteweise in ASCII Symbole umgesetzt

1. Codierung von drei Bytes als vier 6-bit Zeichen



► $3 \times 8\text{-bit} \Leftrightarrow 4 \times 6\text{-bit}$

► 6-bit Binärwerte: $0 \dots 63$

► nutzen 64 (von 128)
7-bit ASCII Symbolen

2. Zahl ASCII Zuordnung der ASCII-Zeichen

0 ... 25	A ... Z
26 ... 51	a ... z
52 ... 61	0 ... 9
62	+
63	/
=	Füllzeichen, falls Anz. Bytes nicht durch 3 teilbar
CR	Zeilenumbruch (opt.), meistens nach 76 Zeichen

Base64-Codierung: Prinzip (cont.)

Text content	M								a								n							
ASCII	77								97								110							
Bit pattern	0	1	0	0	1	1	0	1	0	1	1	0	0	0	0	1	0	1	1	0	1	1	1	0
Index	19								22								5							
Base64-encoded	T								W								F							

- ▶ drei 8-bit Zeichen, neu gruppiert als vier 6-bit Blöcke
- ▶ Zuordnung des jeweiligen Buchstabens/Ziffer
- ▶ ggf. =, == am Ende zum Auffüllen
- ▶ Übertragung dieser Zeichenfolge ist 7-bit kompatibel
- ▶ resultierende Datei ca. 33% größer als das Original

- ▶ im JDK enthalten

Module `java.base`, Package `java.util`: `Base64Encoder`, bzw. `Base64Decoder`

- ▶ diverse andere Packages

- ▶ Apache Commons Codec (commons.apache.org/proper/commons-codec)

`org.apache.commons.codec.binary.Base64InputStream`

`org.apache.commons.codec.binary.Base64OutputStream`

- ▶ JAXB (Java Architecture for XML Binding) in `javax.xml.bind`: `DatatypeConverter`
`parseBase64Binary`, `printBase64Binary`

- ▶ ...

[Jav] Oracle Corporation:

The Java Tutorials – Trail: Internationalization.

Redwood Shores, CA.

`docs.oracle.com/javase/tutorial/i18n`

[Uni] *Unicode - The World Standard for Text and Emoji.*

`home.unicode.org`