64-040 Modul InfB-RS: Rechnerstrukturen

https://tams.informatik.uni-hamburg.de/lectures/2015ws/vorlesung/rs

– Kapitel 5 –

Norman Hendrich

Universität Hamburg
Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik

Technische Aspekte Multimodaler Systeme

Wintersemester 2015/2016

Kapitel 5

Zeichen und Text

Darstellung von Texten

- Ad-Hoc Codierungen
  - Flaggen-Alphabet
  - Braille-Code
  - Morse-Code
- ASCII und ISO-8859-1
- Unicode

Wiederholung: Zeichen

- **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

Wiederholung: Zeichen (cont.)

- **Numerischer Zeichensatz**: Zeichenvorrat aus Ziffern und/oder Sonderzeichen zur Darstellung von Zahlen
- **Alphanumerischer Zeichensatz**: Zeichensatz aus (mindestens) den Dezimalziffern und den Buchstaben des gewöhnlichen Alphabets, meistens auch mit Sonderzeichen (Leerzeichen, Punkt, Komma usw.)

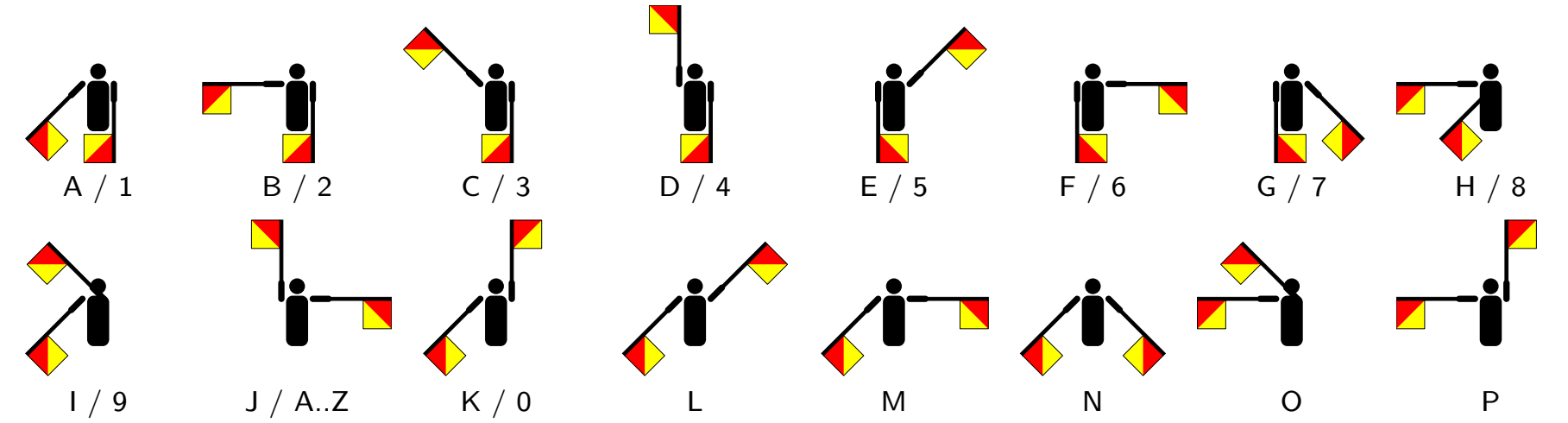
- **Alphabet**: engl. *alphabet*
  Ein in vereinbarter Reihenfolge geordneter Zeichenvorrat

Wiederholung: Zeichen (cont.)

- **Zeichenkette**: engl. *string*
  Eine Folge von Zeichen
- **Wort**: engl. *word*
  Eine Folge von Zeichen, die in einem gegebenen Zusammenhang als Einheit bezeichnet wird
- Worte mit 8 bit werden als **Byte** bezeichnet

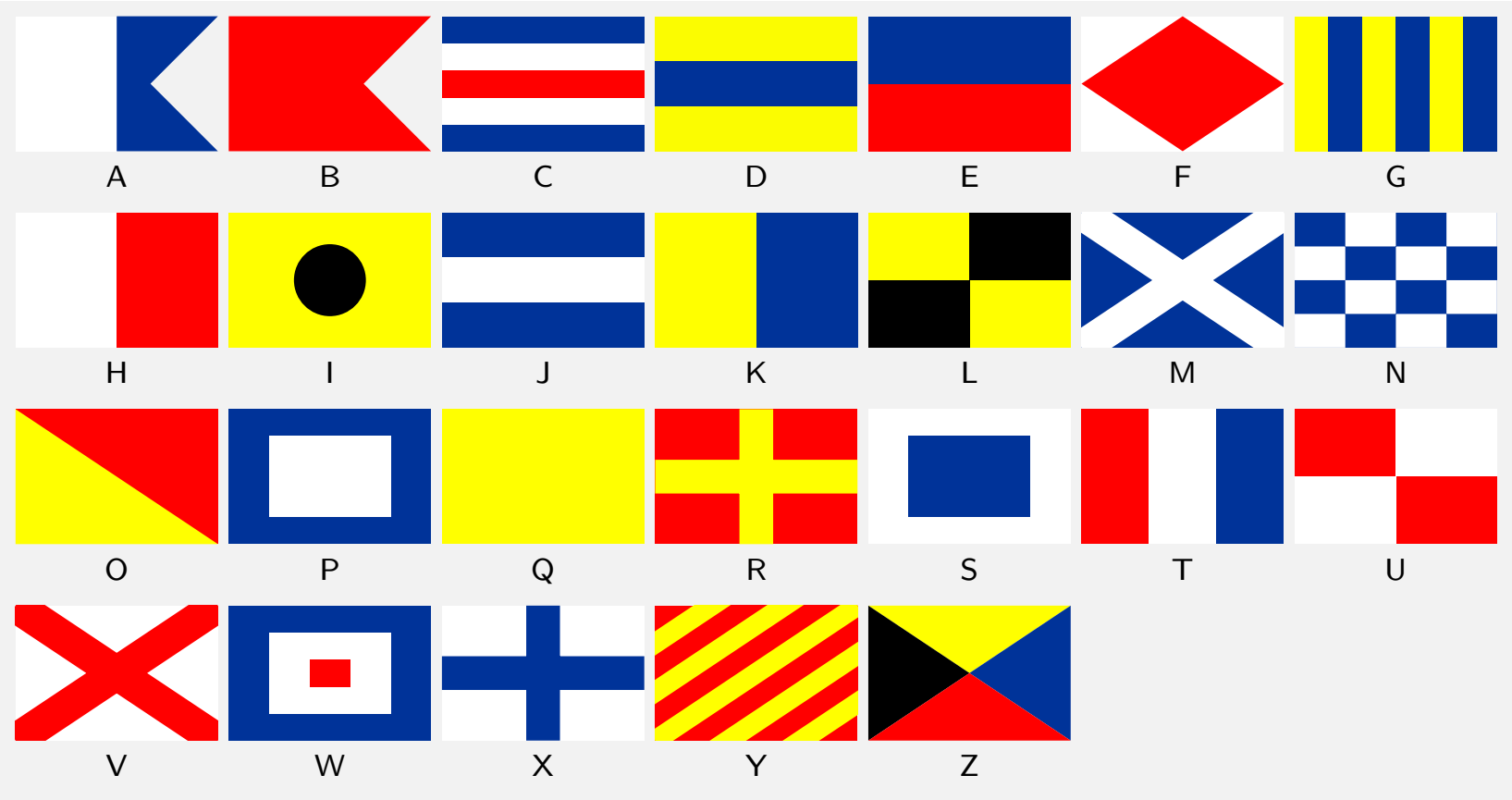
- **Stelle**: engl. *position*
  Die Lage/Position eines Zeichens innerhalb einer Zeichenkette

Flaggen-Signale



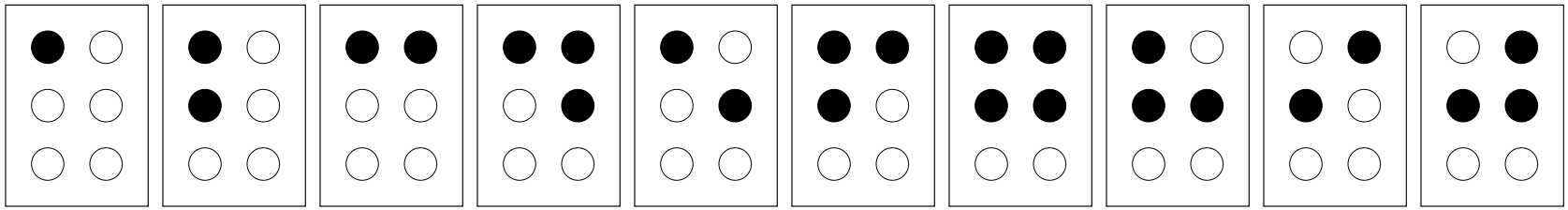
de.wikipedia.org/wiki/Winkeralphabet

Flaggen-Alphabet

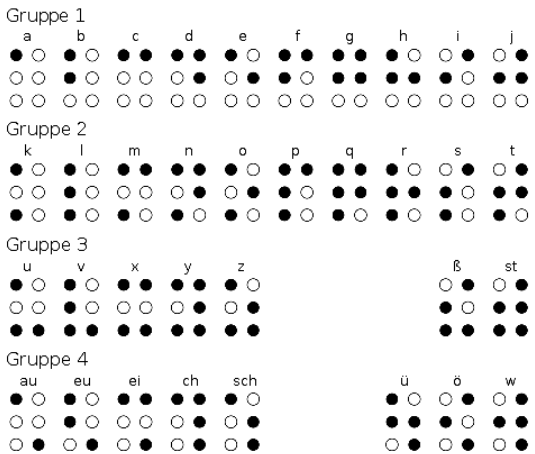


de.wikipedia.org/wiki/Flaggenalphabet

Braille: Blindenschrift



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



Morse-Code

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	••• ––– •••

Morse-Code (cont.)

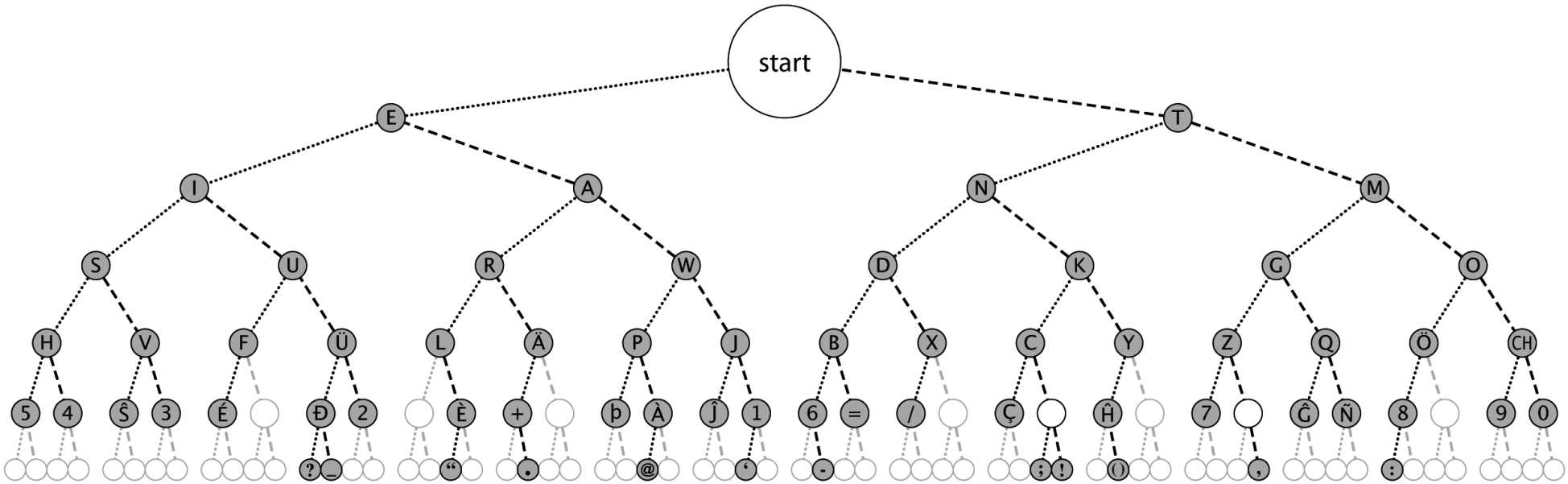
- Eindeutigkeit

Codewort:	● ● ● ● ● – ●
E	●
I	● ●
N	– ●
R	● – ●
S	● ● ●

  - bestimmte Morse-Sequenzen sind mehrdeutig
  - Pause zwischen den Symbolen notwendig

- Codierung
  - Häufigkeit der Buchstaben = 1 / Länge des Codewortes
  - Effizienz: kürzere Codeworte
  - Darstellung als Codebaum

Morse-Code: Baumdarstellung (Ausschnitt)



- Anordnung der Symbole entsprechend ihrer Codierung

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
- 95 druckbare Zeichen: Buchstaben, Ziffern, Sonderzeichen (Codierung im Bereich 21..7E)
- 33 Steuerzeichen (engl: *control characters*) (0..1F,7F)

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

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

http://de.wikipedia.org/wiki/ASCII

ISO-8859 Familie

- 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-8859-1: Codetabelle (1)

Erweiterung von ASCII für westeuropäische Sprachen

Code	…0	…1	…2	…3	…4	…5	…6	…7	…8	…9	…A	…B	…C	…D	…E	…F
0…	*nicht belegt*															
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…	*nicht belegt*															
9…																
A…	*NBSP*	¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	*SHY*	®	¯
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-8859-1: Codetabelle (2)

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-8859-2

Erweiterung von ASCII für slawische Sprachen

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	˘	Ł	¤	Ľ	Ś	§	¨	Š	Ş	Ť	Ź	*SHY*	Ž	Ż
B...	°	ą	˛	ł	´	ľ	ś	ˇ	¸	š	ş	ť	ź	˝	ž	ż
C...	Ŕ	Á	Â	Ă	Ä	Ĺ	Ć	Ç	Č	É	Ę	Ë	Ě	Í	Î	Ď
D...	Đ	Ń	Ň	Ó	Ô	Ő	Ö	×	Ř	Ů	Ú	Ű	Ü	Ý	Ţ	ß
E...	ŕ	á	â	ă	ä	Í	ć	ç	č	é	ę	ë	ě	í	î	ď
F...	đ	ń	ň	ó	ô	ő	ö	÷	ř	ů	ú	ű	ü	ý	ţ	˙

ISO-8859-15

Modifizierte ISO-8859-1 mit € (0xA4)

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*	¡	¢	£	€	¥	Š	§	š	©	ª	«	¬	*SHY*	®	¯
B...	°	±	²	³	Ž	µ	¶	·	ž	¹	º	»	Œ	œ	Ÿ	¿
C...	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D...	Đ	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

Microsoft: Codepages 437, 850, 1252

- Zeichensatz des IBM-PC ab 1981
- Erweiterung von ASCII auf einen 8-bit Code
- einige Umlaute (westeuropäisch)
- Grafiksymbole

- `http://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	å	ç	ê	ë	è	ï	î	ì	Ä	Å
9…	É	æ	Æ	ô	ö	ò	û	ù	ÿ	Ö	Ü	ø	£	Ø	×	ƒ
A…	á	í	ó	ú	ñ	Ñ	ª	º	¿	®	¬	½	¼	¡	«	»
B…	░	▒	▓	│	┤	Á	Â	À	©	╣	║	╗	╝	¢	¥	┐
C…	└	┴	┬	├	─	┼	ã	Ã	╚	╔	╩	╦	╠	═	╬	¤
D…	ð	Ð	Ê	Ë	È	ı	Í	Î	Ï	┘	┌	█	▄	¦	Ì	▀
E…	Ó	ß	Ô	Ò	õ	Õ	µ	þ	Þ	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

- Unterschiedliche Konvention 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	*LF*	`0A`	\n
Windows, DOS, OS/2, CP/M, TOS (Atari)	ASCII	*CR LF*	`0D 0A`	\r\n
Mac OS bis Version 9, Apple II	ASCII	*CR*	`0D`	\r
AIX OS, OS 390	EBCDIC	*NEL*	15	

Unicode: Motivation

- 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:** Umschaltung zwischen mehreren Zeichensätzen durch Spezialbefehle (*Escapesequenzen*).

$\Rightarrow$ **Unicode** als System zur Codierung aller Zeichen aller bekannten (lebenden oder toten) Schriftsysteme

Unicode: Versionen und History

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

- 1991 1.0.0: europäisch, nahöstlich, indisch
- 1992 1.0.1: ostasiatisch (Han)
- 1993 akzeptiert als ISO/IEC-10646 Standard
- ...
- 2014 7.0.0: 113 021 Zeichen = Sprachen, Symbole, Emojis...

`http://www.unicode.org`, `http://de.wikipedia.org/wiki/Unicode`

Unicode: Schreibweise

- 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`

Unicode: in Webseiten (HTML)

- HTML-Header informiert über verwendeten Zeichensatz
- Unterstützung und Darstellung abhängig vom Browser
- Demo `http://www.columbia.edu/~fdc/utf8`

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

Unicode: Demo

http://www.columbia.edu/~fdc/utf8

1. **English:** The quick brown fox jumps over the lazy dog.
2. **Jamaican:** Chruu, a kwik di kwik brong fox a jomp huova di liezi daag de, yu no siit?
3. **Irish:** "An ḃfuil do ċroí ag bualaḋ ó ḟaitíos an ġrá a ṁeall lena ṗóg éada ó ṡlí do leasa ṫú?" "D'ḟuascail Íosa Úrṁac na hÓiġe Beannaiṫe pór Éava agus Áḋaiṁ."
4. **Dutch:** Pa's wijze lynx bezag vroom het fikse aquaduct.
5. **German:** Falsches Üben von Xylophonmusik quält jeden größeren Zwerg. (1)
6. **German:** Im finſteren Jagdſchloß am offenen Felsquellwaſſer patzte der affig-flatterhafte kauzig-höfliche Bäcker über ſeinem verſifften kniffligen C-Xylophon. (2)
7. **Norwegian:** Blåbærsyltetøy ("blueberry jam", includes every extra letter used in Norwegian).
8. **Swedish:** Flygande bäckasiner söka strax hwila på mjuka tuvor.
9. **Icelandic:** Sævör grét áðan því úlpan var ónýt.
10. **Finnish:** (5) Törkylempijävongahdus (This is a perfect pangram, every letter appears only once. Translating it is an art on its own, butI'll say "rude lover's yelp". :-D)
11. **Finnish:** (5) Albert osti fagotin ja töräytti puhkuvan melodian. (Albert bought a bassoon and hooted an impressive melody.)
12. **Finnish:** (5) On sangen hauskaa, että polkupyörä on maanteiden jokapäiväinen ilmiö. (It's pleasantly amusing, that the bicycle is an everyday sight on the roads.)
13. **Polish:** Pchnąć w tę łódź jeża lub osiem skrzyń fig.
14. **Czech:** Příliš žluťoučký kůň úpěl ďábelské kódy.
15. **Slovak:** Starý kôň na hŕbe kníh žuje tíško povädnuté ruže, na stĺpe sa ďateľ učí kvákať novú ódu o živote.
16. **Greek** (monotonic): ξεσκεπάζω την ψυχοφθόρα βδελυγμία
17. **Greek** (polytonic): ξεσκεπάζω τὴν ψυχοφθόρα βδελυγμία
18. **Russian:** Съешь же ещё этих мягких французских булок да выпей чаю.
19. **Russian:** В чащах юга жил-был цитрус? Да, но фальшивый экземпляр! ёъ.
20. **Bulgarian:** Жълтата дюля беше щастлива, че пухът, който цъфна, замръзна като гьон.
21. **Sami (Northern):** Vuol Ruoŧa geđggiid leat máŋga luosa ja čuovžža.
22. **Hungarian:** Árvíztűrő tükörfúrógép.
23. **Spanish:** El pingüino Wenceslao hizo kilómetros bajo exhaustiva lluvia y frío, añoraba a su querido cachorro.
24. **Portuguese:** O próximo vôo à noite sobre o Atlântico, põe freqüentemente o único médico. (3)
25. **French:** Les naïfs ægithales hâtifs pondant à Noël où il gèle sont sûrs d'être déçus en voyant leurs drôles d'œufs abîmés.
26. **Esperanto:** Eĥoŝanĝo ĉiuĵaŭde.
27. **Hebrew:** .דה כיף סתם לשמוע איך תנצח קרפד עץ טוב בגן
28. **Japanese** (Hiragana):

    いろはにほへど　ちりぬるを
    わがよたれぞ　つねならむ
    うゐのおくやま　けふこえて
    あさきゆめみじ　ゑひもせず (4)

Unicode: Demo (cont.)

http://www.columbia.edu/~fdc/utf8

Šota Rustaveli's Veṗxis Ṭq̇aosani, Ȟh, *The Knight in the Tiger's Skin* (Georgian):

> ვეპხის ტყაოსანი შოთა რუსთაველი
>
> ღმერთსი შემვედრე, ნუთუ კვლა დამხსნას სოფლისა შრომასა, ცეცხლს, წყალსა და მიწასა, ჰაერთა თანა მრომასა; მომცნეს ფრთენი და აღვფრინდე, მივჰხვდე მას ჩემსა ნდომასა, დღისით და ღამით ვჰხედვიდე მზისა ელვათა კრთომაასა.

Tamil poetry of Subramaniya Bharathiyar: **சுப்ரமணிய பாரதியார்** (1882-1921):

> **யாமறிந்த மொழிகளிலே தமிழ்மொழி போல் இனிதாவது எங்கும் காணோம்,**
> **பாமரராய் விலங்குகளாய், உலகனைத்தும் இகழ்ச்சிசொலப் பான்மை கெட்டு,**
> **நாமமது தமிழரெனக் கொண்டு இங்கு வாழ்ந்திடுதல் நன்றோ? சொல்லீர்!**
> **தேமதுரத் தமிழோசை உலகமெலாம் பரவும்வகை செய்தல் வேண்டும்.**

Unicode: Latin-Zeichen

- Zeichen im Bereich U+0000 bis U+007F wie ASCII
  www.unicode.org/charts/PDF/U0000.pdf
- Bereich von U+0100 bis U+017F für Latin-A
  Europäische Umlaute und Sonderzeichen
  www.unicode.org/charts/PDF/U0100.pdf
- viele weitere Sonderzeichen ab U+0180
  Latin-B, Latin-C, usw.

Unicode: Mathematische Symbole und Operatoren

Vielfältige Auswahl von Symbolen 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`
- ...

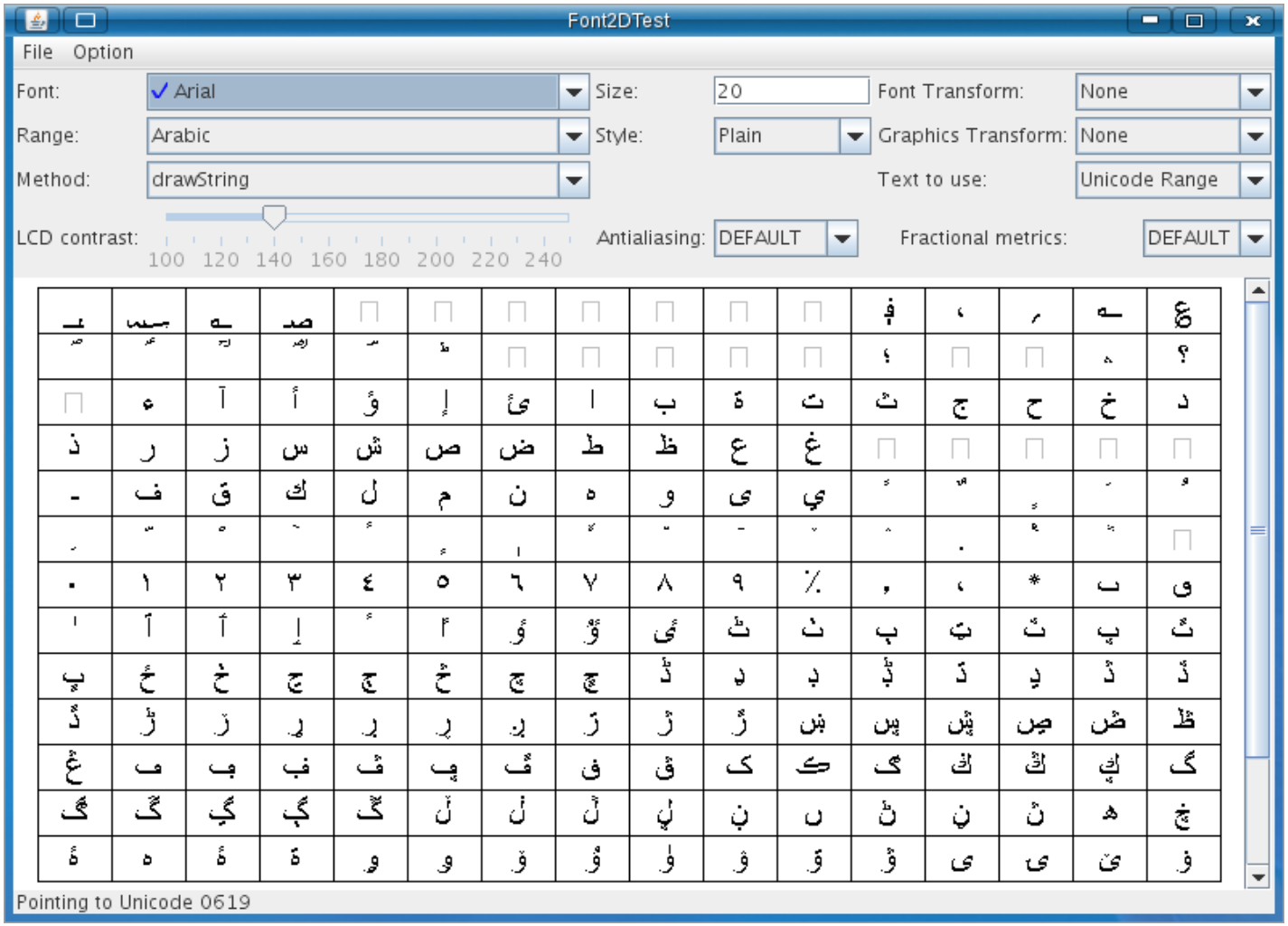
- Dingbats `www.unicode.org/charts/PDF/U2700.pdf`

Unicode: Asiatische Sprachen

Chinesisch (traditional/simplified), Japanisch, Koreanisch

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

Unicode: Java2D Fontviewer



Oracle [JavaD]
JDK demos and samples
`../demo/jfc/Font2DTest`

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: mehrere *Planes* zu je 65 536 Zeichen
- direkte Repräsentation aller Zeichen erfordert 32-bit/Zeichen
- vierfacher Speicherbedarf gegenüber ISO-8859-1

- bei Dateien ist möglichst kleine Dateigröße wichtig
- effizientere Codierung üblich: UTF-16 und UTF-8

UTF-8

Zeichen	Unicode	Unicode binär	UTF-8 binär	UTF-8 hexadezimal
Buchstabe *y*	U+0079	00000000 01111001	01111001	0x79
Buchstabe *ä*	U+00E4	00000000 11100100	11000011 10100100	0xC3 0xA4
Zeichen für eingetragene Marke ®	U+00AE	00000000 10101110	11000010 10101110	0xC2 0xAE
Eurozeichen €	U+20AC	00100000 10101100	11100010 10000010 10101100	0xE2 0x82 0xAC
Violinschlüssel 𝄞	U+1D11E	00000001 11010001 00011110	11110000 10011101 10000100 10011110	0xF0 0x9D 0x84 0x9E

http://de.wikipedia.org/wiki/UTF-8

- effiziente Codierung von „westlichen“ Unicode-Texten
- Zeichen werden mit variabler Länge codiert, 1..4-Bytes
- volle Kompatibilität mit ASCII

UTF-8: Algorithmus

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

Sprach-Einstellungen: Locale

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 ASCII-Texten (z.B. Programm-Quelltexte) zwischen DOS/Windows und Unix/Linux Maschinen

- Umwandeln von `a.txt` in Ausgabedatei `b.txt`:

```
dos2unix -c ascii -n a.txt b.txt
dos2unix -c iso   -n a.txt b.txt
dos2unix -c mac   -n a.txt b.txt
```

- Umwandeln von Unix nach DOS/Windows, Codepage 850:

```
unix2dos -850     -n a.txt b.txt
```

iconv

Das Schweizer-Messer zur Umwandlung von Textcodierungen.

- Optionen
  - `-f, --from-code=<encoding>` Codierung der Eingabedatei
  - `-t, --to-code=<encoding>` Codierung der Ausgabedatei
  - `-l, --list` Liste der unterstützten Codierungen ausgeben
  - `-o, --output=<filename>` Name der Ausgabedatei

- Beispiel

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

base64-Codierung

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

$\Rightarrow$ Verfahren zur Umcodierung der Datei in 7-bit ASCII notwendig

$\Rightarrow$ etabliert ist das **base-64** Verfahren (RFC 2045)

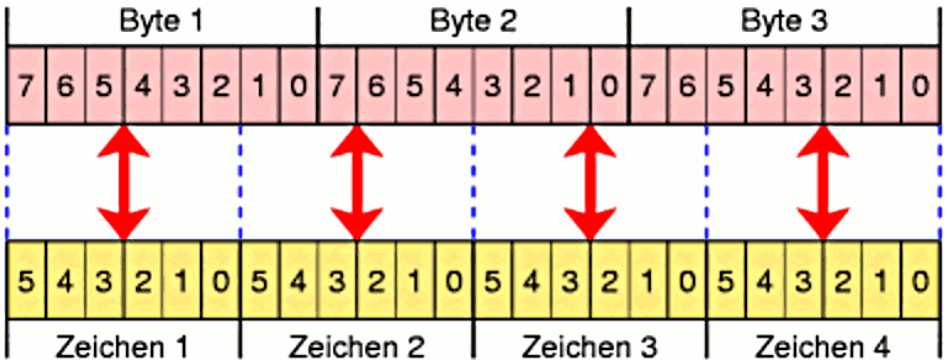
- alle e-mail Dateianhänge und 8-bit Textdateien
- Umcodierung benutzt nur Buchstaben, Ziffern und drei Sonderzeichen

base64-Codierung: Prinzip

- Codierung von drei 8-bit Bytes als vier 6-bit Zeichen
- erfordert 64 der verfügbaren 128 7-bit ASCII Symbole

- Codierung

`A..Z`	Codes:	0..25
`a..z`	Codes:	26..51
`0..9`	Codes:	52..61
`+`	Code:	62
`/`	Code:	63



`=`	Füllzeichen, falls Anzahl der Bytes nicht durch 3 teilbar
`CR`	Zeilenumbruch (optional), 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				46							
Base64-encoded	**T**								**W**				**F**				**u**							

- 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

base64-Codierung: Tools

- im Java JDK enthalten
  aber im inoffiziellen internen Teil
  `sun.misc.BASE64Encoder`, bzw. `sun.misc.BASE64Decoder`

- aber diverse (open-source) Implementierungen verfügbar
  Beispiel: Apache Commons `http://commons.apache.org/codec`
  `org.apache.commons.codec.binary.Base64`
  `org.apache.commons.codec.binary.Base64InputStream`
  `org.apache.commons.codec.binary.Base64OutputStream`

base64-Codierung: Beispiel

openbook.galileodesign.de/javainsel/javainsel_04_010.html

```
import java.io.IOException;
import java.util.*;
import sun.misc.*;

public class Base64Demo
{
  public static void main( String[] args ) throws IOException
  {
    byte[] bytes1 = new byte[ 112 ];
    new Random().nextBytes( bytes1 );

    // buf in String
    String s = new BASE64Encoder().encode( bytes1 );
    System.out.println( s );

    // Zum Beispiel:
    // QFgwDyiQ28/4GsF75fqLMj/bAIWNwOuBmE/SCl3H2XQFpSsSz0jtyR0LU+kLiwWsnSUZljJr97Hy
    // LA3YUbf96Ym2zx9F9Y1N7P5lsOCb/vr2crTQ/gXs757qaJF9E3szMN+E0CSSslDrrzcNBrlcQg==
    // String in byte[]
    byte[] bytes2 = new BASE64Decoder().decodeBuffer( s );
    System.out.println( Arrays.equals(bytes1, bytes2) );    // true
  }
}
```

Literatur

[Uni] The Unicode Consortium; Mountain View, CA.
`www.unicode.org`

[JavaI] Oracle Corporation; Redwood Shores, CA.
The Java Tutorials – Trail: Internationalization.
`docs.oracle.com/javase/tutorial/i18n`

[JavaD] Oracle Corporation: *Java SE Downloads*.
`www.oracle.com/technetwork/java/javase/downloads`

[Ull14] C. Ullenboom: *Java ist auch eine Insel – Einführung, Ausbildung, Praxis*. 11. Auflage, Galileo Press GmbH, 2014.
ISBN 978–3–8362–2873–2
10. Auflage unter `openbook.galileo-press.de/javainsel`