Unicode

Why it became the standard and what the benefits are

Thomas Dauner

Agenda

– Characters in general

– History of charsets

– Unicode

– Unicode's influence on the IT world

– Conclusion

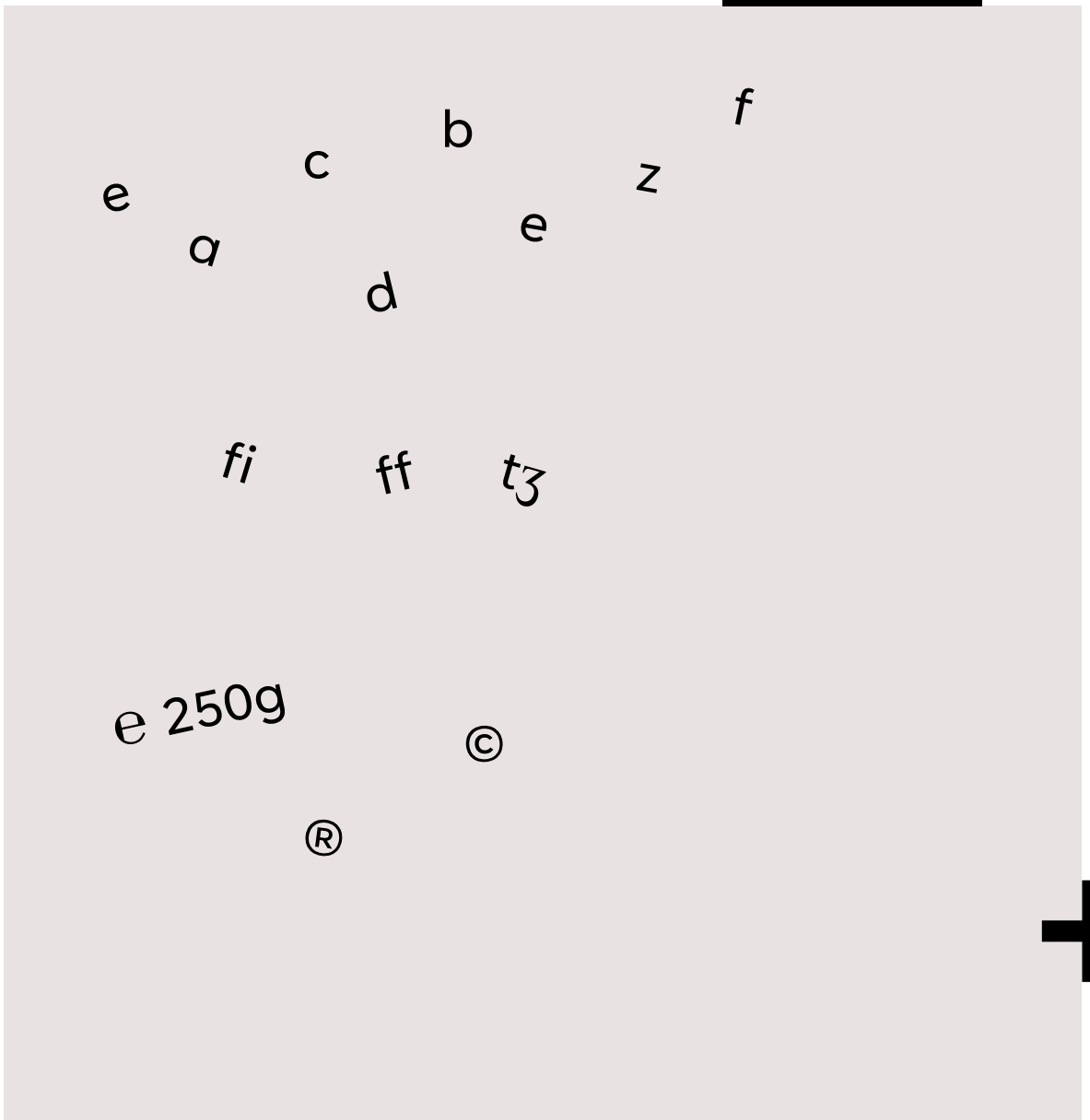
Characters in general

Definitions

– Character = atomic unit of text

– Glyph = one or more combined characters

– Signs may be treated as characters



Characters in general

Technical

– Abstract concept for work on computer needed

– In computers characters are stored as numeric values

– Visual appearance not stored/transmitted

– Systems interpret numeric value & display character

– Simplify interpretation: character collections

Characters in general

Charsets

- Each character has unique numeric value
- Same numeric values in different charsets
- Interpretation = numeric value & charset
  - US-ASCII: 196 = Ä
  - ISO 8859-7 (Greek): 196 = Δ

History of charsets

ASCII

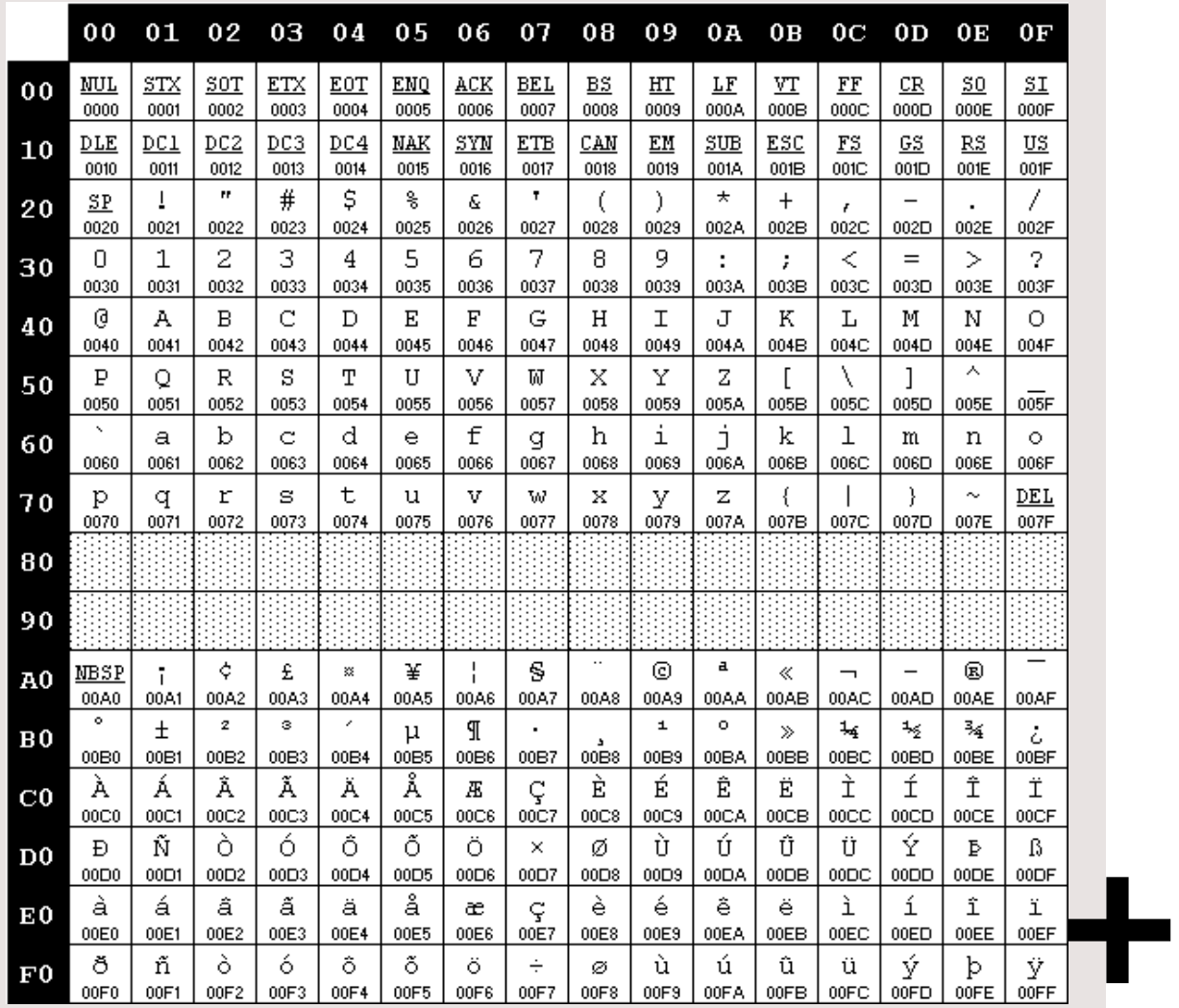
- 7-bit per byte
- 1 byte per character
- 128 characters referenced
  - 33 control characters
  - 95 text characters

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	NUL 0000	STX 0001	SOT 0002	ETX 0003	EOT 0004	ENQ 0005	ACK 0006	BEL 0007	BS 0008	HT 0009	LF 000A	VT 000B	FF 000C	CR 000D	SO 000E	SI 000F
10	DLE 0010	DC1 0011	DC2 0012	DC3 0013	DC4 0014	NAK 0015	SYN 0016	ETB 0017	CAN 0018	EM 0019	SUB 001A	ESC 001B	FS 001C	GS 001D	RS 001E	US 001F
20	SP 0020	! 0021	" 0022	# 0023	$ 0024	% 0025	& 0026	' 0027	(0028	) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
30	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
40	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
50	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C	] 005D	^ 005E	_ 005F
60	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
70	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	| 007C	} 007D	~ 007E	DEL 007F

History of charsets

ISO 8859-1 (ISO Latin 1)

- 8-bit per byte
- 1 byte per character
- 256 characters referenced
- Contains ASCII set + special characters



	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	NUL 0000	STX 0001	SOT 0002	ETX 0003	EOT 0004	ENQ 0005	ACK 0006	BEL 0007	BS 0008	HT 0009	LF 000A	VT 000B	FF 000C	CR 000D	SO 000E	SI 000F
10	DLE 0010	DC1 0011	DC2 0012	DC3 0013	DC4 0014	NAK 0015	SYN 0016	ETB 0017	CAN 0018	EM 0019	SUB 001A	ESC 001B	FS 001C	GS 001D	RS 001E	US 001F
20	SP 0020	! 0021	" 0022	# 0023	$ 0024	% 0025	& 0026	' 0027	(0028	) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
30	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
40	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
50	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C	] 005D	^ 005E	_ 005F
60	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
70	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	| 007C	} 007D	~ 007E	DEL 007F
80																
90																
A0	NBSP 00A0	¡ 00A1	¢ 00A2	£ 00A3	¤ 00A4	¥ 00A5	¦ 00A6	§ 00A7	¨ 00A8	© 00A9	ª 00AA	« 00AB	¬ 00AC	- 00AD	® 00AE	¯ 00AF
B0	° 00B0	± 00B1	² 00B2	³ 00B3	´ 00B4	µ 00B5	¶ 00B6	· 00B7	¸ 00B8	¹ 00B9	º 00BA	» 00BB	¼ 00BC	½ 00BD	¾ 00BE	¿ 00BF
C0	À 00C0	Á 00C1	Â 00C2	Ã 00C3	Ä 00C4	Å 00C5	Æ 00C6	Ç 00C7	È 00C8	É 00C9	Ê 00CA	Ë 00CB	Ì 00CC	Í 00CD	Î 00CE	Ï 00CF
D0	Ð 00D0	Ñ 00D1	Ò 00D2	Ó 00D3	Ô 00D4	Õ 00D5	Ö 00D6	× 00D7	Ø 00D8	Ù 00D9	Ú 00DA	Û 00DB	Ü 00DC	Ý 00DD	Þ 00DE	ß 00DF
E0	à 00E0	á 00E1	â 00E2	ã 00E3	ä 00E4	å 00E5	æ 00E6	ç 00E7	è 00E8	é 00E9	ê 00EA	ë 00EB	ì 00EC	í 00ED	î 00EE	ï 00EF
F0	ð 00F0	ñ 00F1	ò 00F2	ó 00F3	ô 00F4	õ 00F5	ö 00F6	÷ 00F7	ø 00F8	ù 00F9	ú 00FA	û 00FB	ü 00FC	ý 00FD	þ 00FE	ÿ 00FF

History of charsets

Conclusion

– First charsets developed in USA

– Expanded by regional charsets

– Referencing of characters in different Charsets not unique

– Program which decodes text needs correct charset

– Decoding with wrong charset = error/wrong text

Unicode

Overview

– Tries to solve charset jungle

– Unites all charsets and characters in one charset

– Every character has unique identifier

– Using up to 4 bytes per character

Unicode

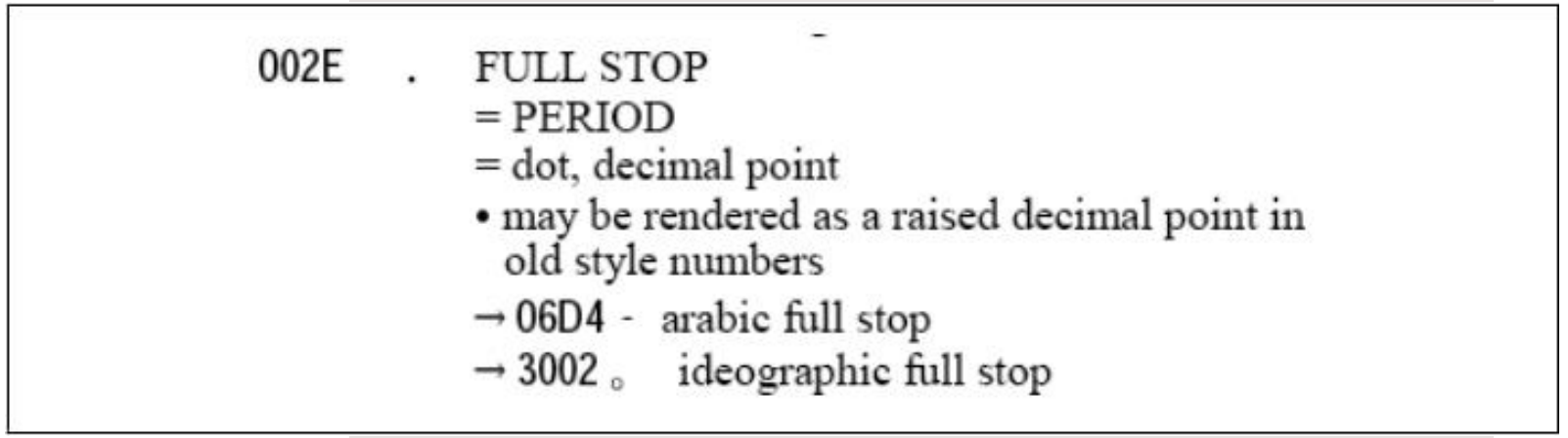
10 Design Principles

- Universality
- Efficiency
- Characters, not glyphs
- Semantics
- Plain text
- Unification
- Dynamic composition
- Logical order
- Equivalent sequences
- Convertibility

Unicode

Character definition

- Mandatory:
  - Unicode number
  - Representative glyph
  - Unicode name
- Optional
  - Old names
  - Comments
  - Cross references



Unicode

Unicode encodings: UTF-32

- Fixed 4 bytes per character
  - 21 bits used on maximum
  - 11 bits per character space wasted
- Fast data access (base+4*(n-1))

Unicode

Unicode encodings: UTF-16

- 2 bytes per character
- If needed: identifier split into 2 16-bit parts
  - Part 1: High value
  - Part 2: Low value
- No direct data access possible

Unicode

Unicode encodings: UTF-8

- 1 byte per character
- If needed: 2, 3 or 4 bytes per character
- No direct data access possible

Code number in binary	Octet 1	Octet 2	Octet 3	Octet 4
00000000 0xxxxxxx	0xxxxxxx			
00000yyy yyxxxxxx	110yyyyy	10xxxxxx		
zzzzyyyy yyxxxxxx	1110zzzz	10yyyyyy	10xxxxxx	
uuuww zzzzyyyy yyxxxxxx	11110uuu	10wwzzzz	10yyyyyy	10xxxxxx

Unicode

Criticism

- Very complex
- Inefficient
- Reasonability of supporting over 100 000 characters

Class of characters	Range of characters	UTF-8	UTF-16	UTF-32
Basic Latin (ASCII)	U+0000 to U+007F	1	2	4
Latin 1 Suppl., …, Thaana	U+0080 to U+07FF	2	2	4
Rest of BMP	U+0800 to U+FFFF	3	2	4
Outside BMP	U+10000 to U+10FFFF	4	4	4

Unicodes influence on the IT world

Improvements

– Multilingual applications

– Multilingual documents (e.g. german and russian)

– Solved different-language-different-charset problem

– Character identifiers no longer ambigous

– Easy charset conversions

– Covers different areas of use (e.g. internet, programming, ...)

Unicodes influence on the IT world

Problems

- OS support Unicode, software doesn‘t
- Support Unicode != display all characters
- MySQL:
  - Utf8 encoding: supports 3 byte characters (!= Unicode UTF8)
  - Utf8-mb4 encoding: supports 4 byte characters (=Unicode UTF8)
- Interapplication communication

Conclusion

– Unicode solved problem of many different small charsets

– Unicode solved problem of ambiguous identifiers

– Unicode provides one charset with all known + used characters

– Different encodings for different usages

– Advantages in multilingual software developments

– Still some problems, will be disappear in future


Time for questions!

References

J. K. Korpela, „Unicode Explained,“ O'Reilly Media, Inc., 2006, pp. 6-10.

J. K. Korpela, „Unicode Explained,“ O'Reilly Media, Inc., 2006, pp. 11-15.

J. K. Korpela, „Unicode Explained,“ O'Reilly Media, Inc., 2006, pp. 3-5.

„Charset.org,“ [Online]. Available: https://www.charset.org/charsets/us-ascii. [Zugriff am 18 12 2022].

„Charset.org,“ [Online]. Available: https://www.charset.org/charsets/iso-8859-7. [Zugriff am 18 12 2022].

R. Gillam, „Unicode Demystified,“ Addison-Wesley Professional, 2003, p. Chapter: What Unicode Is.

J. K. Korpela, „Unicode Explained,“ O'Reilly Media, Inc., 2006, pp. 124-126.

J. K. Korpela, „Unicode Explained,“ O'Reilly Media, Inc., 2006, pp. 127-129.

R. Gillam, „Unicode Demystified,“ Addison-Wesley Professional, 2003, p. Chapter: What Unicode Isn't.

J. K. Korpela, „Unicode Explained,“ O'Reilly Media, Inc., 2006, pp. 157-161.

„unicode.org,“ [Online]. Available: https://www.unicode.org/Public/UCD/latest/charts/CodeCharts.pdf. [Zugriff am 18 12 2022].

„Unicode Standard 15.0.0,“ 13 09 2022. [Online]. Available: https://www.unicode.org/versions/Unicode15.0.0/. [Zugriff am 18 12 2022].

J. K. Korpela, „Unicode Explained,“ O'Reilly Media, Inc., 2006, pp. 16-20.

J. K. Korpela, „Unicode Explained,“ O'Reilly Media, Inc., 2006, pp. 21, Figure 1-2.

J. K. Korpela, „Unicode Explained,“ O'Reilly Media, Inc., 2006, pp. 301-311.

J. K. Korpela, „Unicode Explained,“ O'Reilly Media, Inc., 2006, pp. 306, Table 6-1.

J. K. Korpela, „Unicode Explained,“ O'Reilly Media, Inc., 2006, pp. 203-207.

J. K. Korpela, „Unicode Explained,“ O'Reilly Media, Inc., 2006, pp. 327, Table 6-5.

J. K. Korpela, „Unicode Explained,“ O'Reilly Media, Inc., 2006, pp. 326-329.

MacLemon, „Emoji, wie funktionieren die eigentlich?,“ in *GPN19*, 2019.