

# Encoding – Eine Geschichte voller Missverständnisse

---

Felix Dreißig

5. November 2019

FSI Informatik Lightning Talks #1

```
Folien.tex:36: Package inputenc Error: Unicode  
character ? (U+FFFD)  
Folien.tex:36: leading text: \maketitle
```

## Stand 2019

```
Folien.tex:36: Package inputenc Error: Unicode  
character  $\diamond$  (U+FFFD)
```

```
Folien.tex:36: leading text: \maketitle
```



# Die Anfänge

- Morse, Baudot etc. (19. Jahrhundert)
- BCDIC 1928, ASCII ~ 1960, EBCDIC 1963

```
> man ascii
```

```
      2 3 4 5 6 7          30 40 50 60 70 80 90 100 110 120
-----
0:   0 @ P ` p    0:   ( 2 < F P Z d n x
1:   ! 1 A Q a q  1:   ) 3 = G Q [ e o y
2:   " 2 B R b r  2:   * 4 > H R \ f p z
3:   # 3 C S c s  3:   ! + 5 ? I S ] g q {
4:   $ 4 D T d t  4:   " , 6 @ J T ^ h r |
5:   % 5 E U e u  5:   # - 7 A K U _ i s }
6:   & 6 F V f v  6:   $ . 8 B L V ` j t ~
7:   ' 7 G W g w  7:   % / 9 C M W a k u DEL
8:   ( 8 H X h x  8:   & 0 : D N X b l v
9:   ) 9 I Y i y  9:   ' 1 ; E O Y c m w
A:  * : J Z j z
B:  + ; K [ k {
C:  , < L \ l |
D:  - = M ] m }
E:  . > N ^ n ~
F:  / ? 0 _ o DEL
```

# ISO-8859 (ab 1987)

## Für Westeuropa: ISO-8859-1 (Latin-1)

7_112	p	q	r	s	t	u	v	w	x	y	z	{		}	~	
	0070	0071	0072	0073	0074	0075	0076	0077	0078	0079	007A	007B	007C	007D	007E	
8_128																
9_144																
A_160	NBSP	i	ç	£	¤	¥		§	¨	©	à	«	¬	SHY	®	-
	00A0	00A1	00A2	00A3	00A4	00A5	00A6	00A7	00A8	00A9	00AA	00AB	00AC	00AD	00AE	00AF
B_176	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
	00B0	00B1	00B2	00B3	00B4	00B5	00B6	00B7	00B8	00B9	00BA	00BB	00BC	00BD	00BE	00BF
C_192	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
	00C0	00C1	00C2	00C3	00C4	00C5	00C6	00C7	00C8	00C9	00CA	00CB	00CC	00CD	00CE	00CF
D_208	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
	00D0	00D1	00D2	00D3	00D4	00D5	00D6	00D7	00D8	00D9	00DA	00DB	00DC	00DD	00DE	00DF
E_224	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
	00E0	00E1	00E2	00E3	00E4	00E5	00E6	00E7	00E8	00E9	00EA	00EB	00EC	00ED	00EE	00EF
F_240	ø	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ
	00F0	00F1	00F2	00F3	00F4	00F5	00F6	00F7	00F8	00F9	00FA	00FB	00FC	00FD	00FE	00FF

# Unicode (ab 1991)

- Große Tabelle
- Ursprünglich  $2^{16}$  Zeichen
- Heute 1.114.112 *Code points* in 17 *Planes* à 65.536 ( $2^{16}$ )
- Belegt: Ca. 138.000 *Code points*
- Zusätzlich: *Composite characters*

# Was ist drin im Unicode?

## Plane 0: *Basic Multilingual Plane*

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F	
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F	
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F	
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F	
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F	
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F	
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F	
80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F	
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F	
A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF	
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF	
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF	
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF	
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF	
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF	

- Latin script
- Non-Latin European scripts
- African scripts
- Middle Eastern and Southwest Asian scripts
- South and Central Asian scripts
- Southeast Asian scripts
- East Asian scripts
- CJK characters
- Indonesian and Oceanic scripts
- American scripts
- Notational systems
- Symbols
- Private use
- UTF-16 surrogates
- Unallocated code points

As of Unicode 12.0

## Plane 1: *Supplementary Multilingual Plane*



- UCS-2: 16 Bit fix → Nur *Basic Multilingual Plane*
- UTF-16: Erweiterung von UCS-2, Zeichen außerhalb der BMP werden durch *Surrogate* ( $2 \times 16$  Bit) codiert
- UCS-4: 32 Bit fix



# UTF-8: Geschichte

Subject: UTF-8 history  
From: "Rob 'Commander' Pike" <r (at) google.com>  
Date: Wed, 30 Apr 2003 22:32:32 -0700 (Thu 06:32 BST)  
To: mkuhn (at) acm.org, henry (at) spsystems.net  
Cc: ken (at) entrisphere.com

Looking around at some UTF-8 background, I see the same incorrect story being repeated over and over. The incorrect version is:

1. IBM designed UTF-8.
2. Plan 9 implemented it.

That's not true. UTF-8 was designed, in front of my eyes, on a placemat in a New Jersey diner one night in September or so 1992.

<https://www.cl.cam.ac.uk/~mgk25/ucs/utf-8-history.txt>

## UTF-8: Funktionsweise

Number of bytes	Bits for code point	First code point	Last code point	Byte 1	Byte 2	Byte 3	Byte 4
1	7	U+0000	U+007F	0xxxxxxx			
2	11	U+0080	U+07FF	110xxxxx	10xxxxxx		
3	16	U+0800	U+FFFF	1110xxxx	10xxxxxx	10xxxxxx	
4	21	U+10000	U+10FFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

- Abwärtskompatibel mit ASCII
- Für viele Alphabete relativ effizient
- Selbstsynchronisierend

- Heuristiken

```
> file -i foo.txt
foo.txt: text/plain; charset=utf-8
```

- BOM
- Spezifikation
- Auszeichnung

```
<meta charset="UTF-8" />
```

- Verwendet UTF-8 zum Speichern und Übertragen
- Nutzt intern die Unicode-Strings eurer Programmiersprache
- Baut *Unicode-Sandwiches*

# Encoding – Eine Geschichte voller Missverständnisse

---

Felix Dreißig

5. November 2019

FSI Informatik Lightning Talks #1