XML en Unicode

Marc de Graauw
http://www.marcdegraauw.com/

HTML - een voorbeeld

```
<HTML>
<HEAD>
<TITLE>Marc de Graauw</TITLE>
</HEAD>
<BODY>
<H1>Marc de Graauw</H1>
<P>Geslacht: Man</P>
<H2>Opleidingen</H2>
<OL>
<LI>VWO</LI>
<LI>kandidaats Biologie</LI>
<LI>doctoraal filosofie</LI>
</OL>
</BODY>
</HTML>
```

XML - een voorbeeld

```
<?xml version="1.0" encoding="UTF-8"?>

<persoon>
   <persoonsgegevens geslacht="Man">
      <achternaam>Graauw</achternaam>
      <voorvoegsel>de</voorvoegsel>
      <voornaam>Marc</voornaam>
   </persoonsgegevens>
   <opleiding>VWO</opleiding>
   <opleiding>kandidaats Biologie</opleiding>
   <opleiding>doctoraal filosofie</opleiding>
</persoon>
```

Vóór Unicode

bits	0110 1101	0110 0001	0111 0010	0110 0011
bytes dec	109	97	114	99
bytes hex	6d	61	72	63
ascii	m	a	r	c

- 7 bits, 128 tekens, 95 afdrukbaar
- Engels: cijfers, letters, leestekens
- niet: Frans: ê ç Duits: ß ä Grieks: ε Ω Nederlands: ë ï
- Spaans, Arabisch, Fins, Russisch, Chinees, Thais, etc. etc.

Vóór Unicode

bytes 0 - 127

ASCII

```
 !"#$%&'()*+,-./
0123456789:;<=>?
@ABCDEFGHIJKLMNO
PQRSTUVWXYZ[\]^_
`abcdefghijklmno
pqrstuvwxyz{|}~
```

bytes 128 - 255

code page 473

Engels, meeste Frans, Duits IBM PC

code page 850

West Europees + Á ß

DEC Multi-national Character Set

Windows 1252

ISO-8859-1 (ISO-Latin)

West-Europees

ISO-8859-15

ISO-Latin + ‘IJ’sland + ‘œ’uf

ISO-Latin + œ, €

Shift-JIS

Japans

EBCDIC

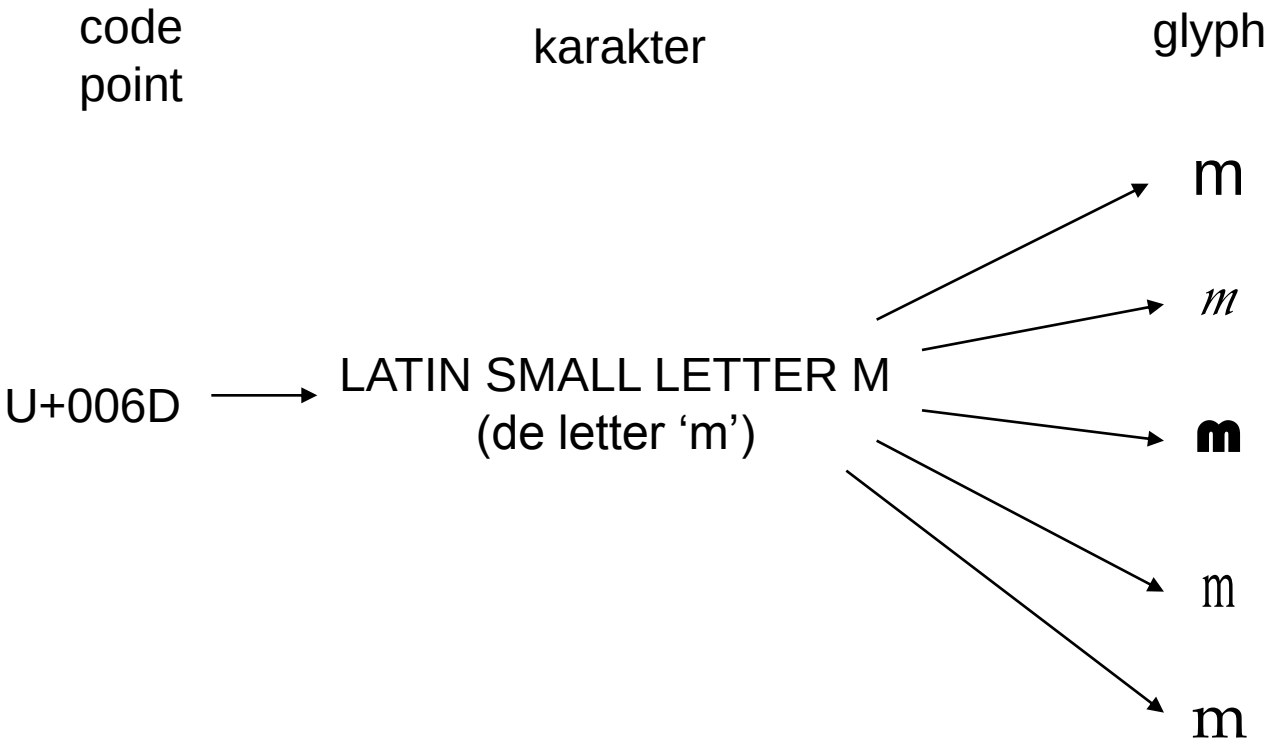
IBM mainframe

etc. etc. etc....

Vóór Unicode

	CP 437	CP 850	Windows 1252	ISO-Latin-1	ISO-Latin-15	Unicode
a	61	61	61	61	61	0061
á	a0	a0	e1	e1	e1	00e1
ä	84	84	e4	e4	e4	00e4
€	-	-	80	-	a4	20ac
õ	-	e4	f5	f5	f5	00f5
Ç	80	80	c7	c7	c7	00c7
œ	-	-	9c	-	bd	0153

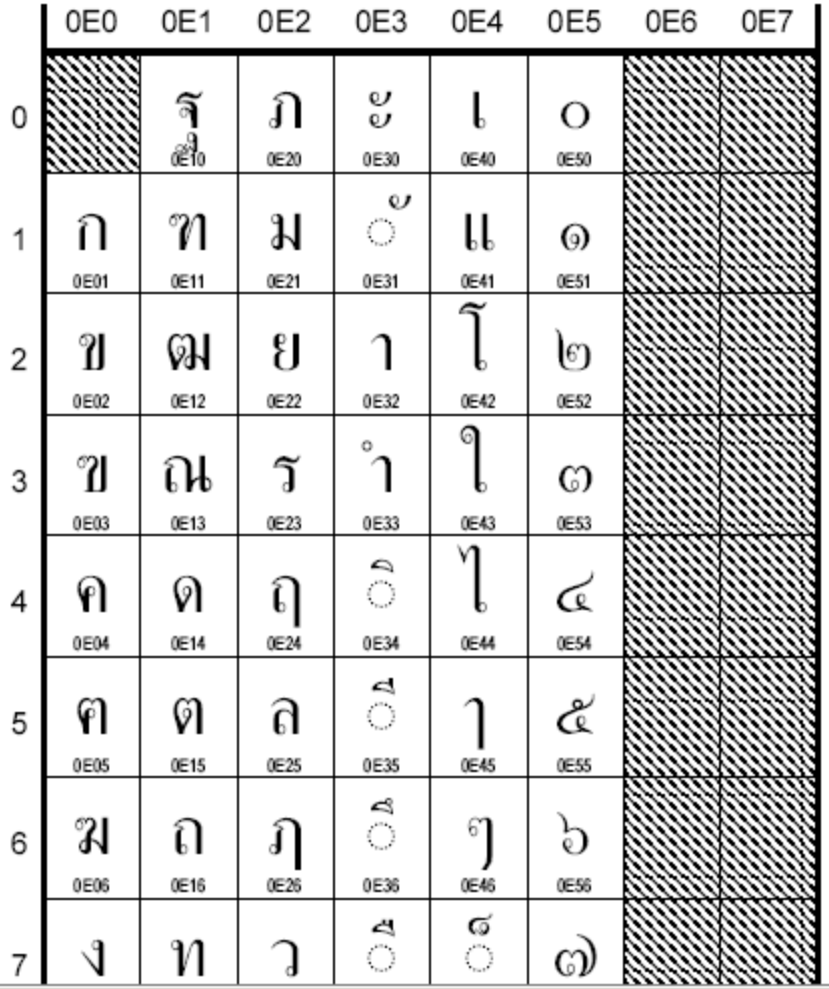
Unicode



Marc de Graauw
http://www.marcdegraauw.com/

	000	001	002	003	004	005	006	007
0	NUL 0000	DLE 0010	SP 0020	0 0030	@ 0040	P 0050	` 0060	p 0070
1	SOH 0001	DC1 0011	! 0021	1 0031	A 0041	Q 0051	a 0061	q 0071
2	STX 0002	DC2 0012	" 0022	2 0032	B 0042	R 0052	b 0062	r 0072
3	ETX 0003	DC3 0013	# 0023	3 0033	C 0043	S 0053	c 0063	s 0073
4	EOT 0004	DC4 0014	$ 0024	4 0034	D 0044	T 0054	d 0064	t 0074
5	ENQ 0005	NAK 0015	% 0025	5 0035	E 0045	U 0055	e 0065	u 0075
6	ACK 0006	SYN 0016	& 0026	6 0036	F 0046	V 0056	f 0066	v 0076
7	BEL	ETB	'	7	G	W	g	w

0E00 Thai 0E7F

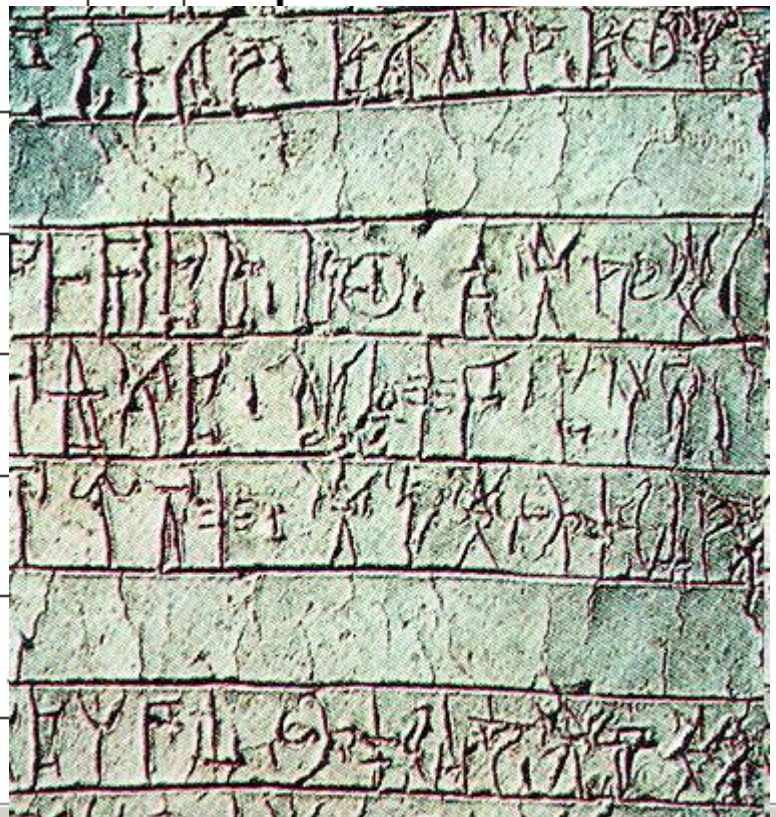


4E00 **CJK Unified Ideographs** **4EFF**

	4E0	4E1	4E2	4E3	4E4	4E5	4E6	4E7	4E8	4E9	4EA	4EB	4EC	4ED	4EE	4EF
0	一 4E00	丐 4E10	北 4E20	丰 4E30	乀 4E40	乐 4E50	习 4E60	买 4E70	亀 4E80	亏 4E90	亠 4EA0	京 4EB0	什 4EC0	仐 4ED0	仠 4EE0	仰 4EF0
1	丁 4E01	丑 4E11	两 4E21	卯 4E31	乁 4E41	永 4E51	乡 4E61	乱 4E71	乾 4E81	云 4E91	亡 4EA1	佰 4EB1	仁 4EC1	仑 4ED1	仡 4EE1	仱 4EF1
2	丂 4E02	刃 4E12	丢 4E22	串 4E32	义 4E42	乒 4E52	乜 4E62	乲 4E72	亂 4E82	互 4E92	亢 4EA2	亲 4EB2	仂 4EC2	仒 4ED2	仢 4EE2	仲 4EF2
3	七 4E03	专 4E13	丣 4E23	弗 4E33	乃 4E43	乓 4E53	乩 4E63	乳 4E73	亃 4E83	亓 4E93	亣 4EA3	亳 4EB3	仃 4EC3	仓 4ED3	代 4EE3	仳 4EF3
4	丄 4E04	且 4E14	两 4E24	临 4E34	乄 4E44	乔 4E54	乣 4E64	乴 4E74	亄 4E84	五 4E94	交 4EA4	亴 4EB4	仄 4EC4	仔 4ED4	令 4EE4	仴 4EF4
5	丅 4E05	丕 4E15	严 4E25	丵 4E35	久 4E45	乕 4E55	乥 4E65	乵 4E75	亅 4E85	井 4E95	亥 4EA5	亵 4EB5	仅 4EC5	仕 4ED5	以 4EE5	仵 4EF5
6	丆 4E06	世 4E16	並 4E26	丶 4E36	乆 4E46	乖 4E56	书 4E66	乶 4E76	了 4E86	三 4E96	亦 4EA6	亶 4EB6	仆 4EC6	他 4ED6	仦 4EE6	件 4EF6
7	万	丗	丧	丷	乇	乗	乧	乷	冂	岂	产	廉	仇	仗	夫	价

10080 Linear B Ideograms 100FF

	1008	1009	100A	100B	100C	100D	100E	100F
0	10080	10090	100A0	100B0	100C0	100D0	100E0	100F0
1	10081	10091	100A1	100B1	100C1			
2	10082	10092	100A2	100B2	100C2			
3	10083	10093	100A3	100B3	100C3			
4	10084	10094	100A4	100B4	100C4			
5	10085	10095	100A5	100B5	100C5			
6	10086	10096	100A6	100B6	100C6			
7								



12000 Cuneiform 120FF

	1200	1201	1202	1203	1204	1205	1206	1207	1208	1209	120A	120B	120C	120D	120E	120F
0	12000	12010	12020	12030	12040	12050	12060	12070	12080	12090	120A0	120B0	120C0	120D0	120E0	120F0
1	12001	12011	12021	12031	12041	12051	12061	12071	12081	12091	120A1	120B1	120C1	120D1	120E1	120F1
2	12002	12012	12022	12032	12042	12052	12062	12072								
3	12003	12013	12023	12033	12043	12053	12063	12073								
4	12004	12014	12024	12034	12044	12054	12064	12074								
5	12005	12015	12025	12035	12045	12055	12065	12075								
6	12006	12016	12026	12036	12046	12056	12066	12076								
7																

http://www.evertype.com/standards/csur/tengwar.html

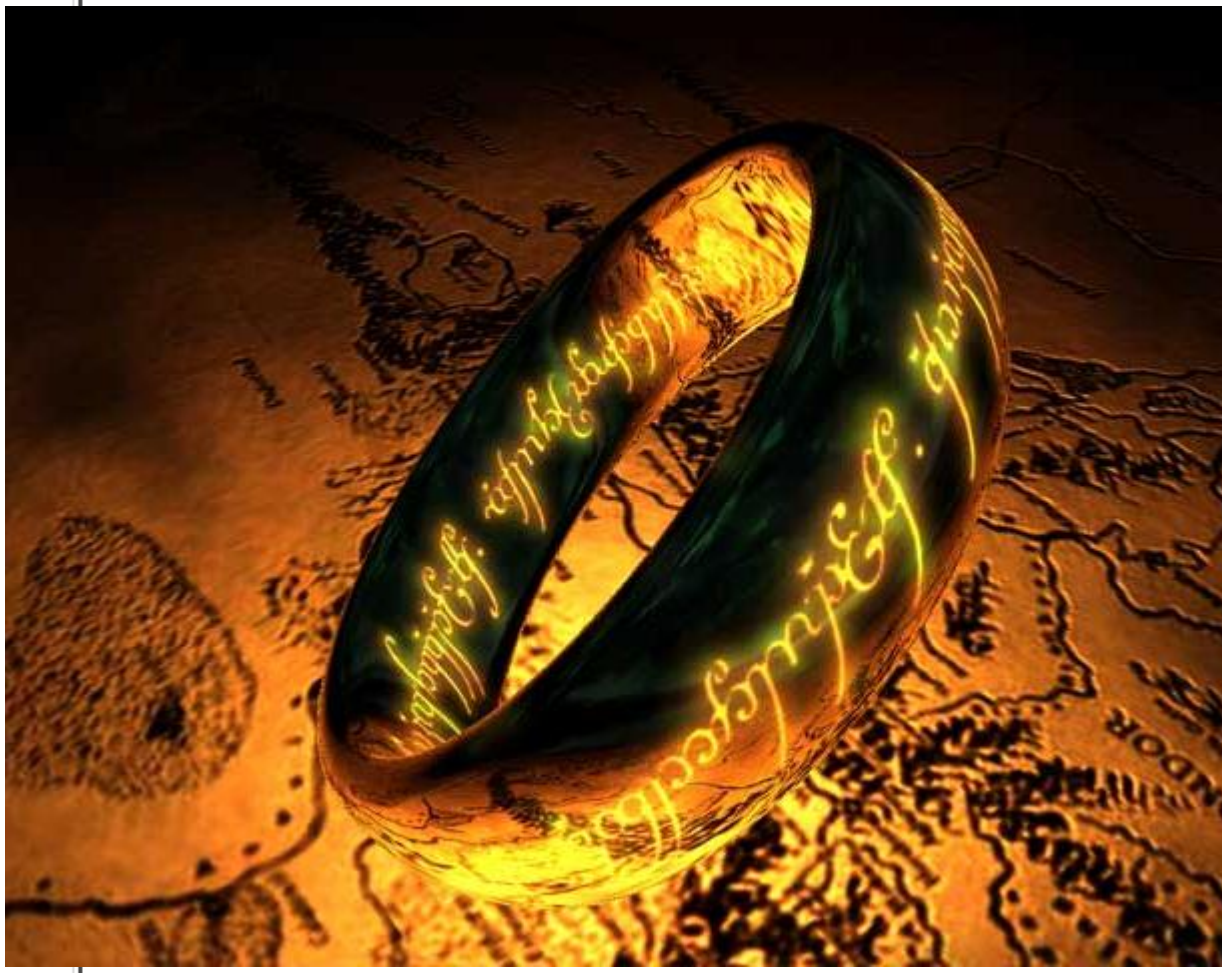
klingon unicode

klingon unicode

le Calendar | Unicode - Wikipedia, the free encyclop... | ConScript Unicode Registry

E00	E01	E02	E03	E04	E05	E06	E07
p	b	ʎ	ɔ	◌̈	·		
p	b	ɗ	ɔ	◌̤	:		
q	ɑ	ʎ	c	◌̈	ʃ	ɐ	
q	ɑ	o	c	◌̤	β	ɑ	
b	ɒ	l		◌̈	~	l	
b	ɒ	ı		◌̤	~	ɾ	
ɑ	ɑ	l		◌́	◌̤	ɾ	
ɑ	ɑ	ɑ		◌́	◌̤	ɹ	
b	ɒ	þ		◌̈	◌̤	þ	
b	ɒ	þ		◌̈		b	
ɗ	ᑌ	ɗ		◌̈	◌̤	ɔ	
ɗ	ᑎ	ɗ		◌̈		ɔ	
b	ɔ	b		◌̄	◌̤	◌̇	
b	ɔ	b		◌̄	◌̤	◌̇	

Tengwar – Tolkien - niet officieel (private use range)



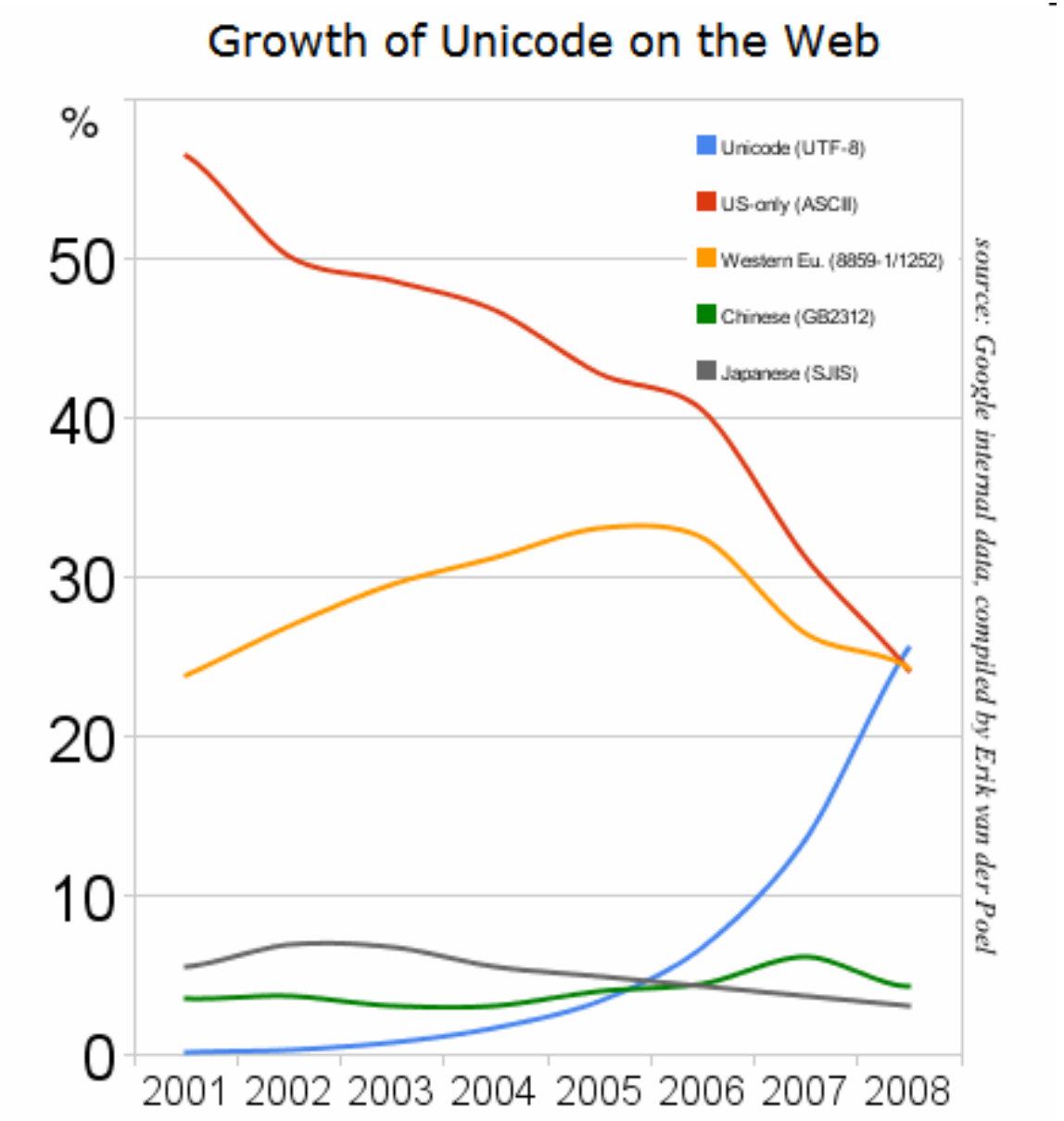
Unicode encodings

- Unicode
  - U+006D = ‘m’
  - karakter 0 – 255: gelijk aan ISO-Latin-1
  - 1.114.112 code points (0 – 10FFFF)
- UTF-16 encoding
  - 4 bytes
  - 0000 – FFFF: gelijk aan Unicode nummer
  - Byte Order Mark
    - U+FEFF (ZERO-WIDTH NO-BREAK SPACE)
    - byte-swapped = U+FFFE = geen legaal karakter
  - efficiënt voor Chinees en Japans
- UTF-8
  - 1 tot 4 bytes
  - 0 – 127: gelijk aan ASCII
  - ergo: ASCII tekst is altijd ook UTF-8 tekst
  - efficiënt voor Westerse talen

Unicode encodings

teken	code point	UTF-8	UTF-16	ISO-Latin-1	ASCII
spatie	U+0020	20	00 20	20	20
a	U+0061	61	00 61	61	61
ä	U+00E4	C3 A4	00 E4	E4	-

- XML
- Java
- .NET



Marc de Graauw
http://www.marcdegraauw.com/