Het invoeren en afdrukken van de Latin-1 (ISO-8859-1) Characterset

Ton Biegstraaten

Samenvatting

Ten behoeve van een collegedictaat is een tabel nodig met de Latin-1 of ISO-8859-1 characterset. Dit artikel behandelt het invoeren en afdrukken van deze characterset met behulp van LaTeX2e.

1 Enige achtergrond informatie

1.1 TeX

Sinds TeX versie 3 kunnen naast de gebruikelijke charactersets met 128 tekens ook sets met 256 tekens gedefinieerd en gebruikt worden. Een voorbeeld is de ISO-8859-1 of Latin-1 characterset welke voldoende symbolen bevat om de meeste europese talen te kunnen verwerken.

1.2 Extended Computer Modern fonts

In 1990 gedurende de TUG conferentie in Cork, Ierland is een nieuw font coderingsschema vastgesteld welke veel europese talen met een latijns schrift ondersteunt. De eerste 128 posities van dit schema komen overeen met de bestaande codering voor de computer modern fonts. De laatste 128 posities worden ingenomen door characters met accenten en bijzondere taalspecifieke characters. Voorheen konden sommige van deze characters alleen als combinatie van andere characters worden afgedrukt. Anderen konden niet worden afgedrukt.

Omdat verschillende fonts verschillende coderingsschema's kunnen hebben, moeten ze onderscheiden kunnen worden. Ze krijgen daarom een unieke naam. De nieuwe codering heeft de naam: `T1`, de oude `OT1`, old of obsolete `T1`. Zo zijn er meerdere coderingsschema's gedefinieerd.

Aan een codering alleen heb je niet veel, er horen fonts bij waardoor de codering te gebruiken is. Deze fonts zijn begin 1997 opgeleverd en worden 'extended computer modern fonts' of `ec` fonts genoemd. Tijdens de ontwikkeling zijn ze beschikbaar geweest als `dc` fonts. Alle voorbeelden in dit artikel werken zowel met de `dc` als met de `ec` fonts.

Naast de `ec` fonts was er behoefte aan fonts met symbolen die voorheen alleen in mathematische fonts beschikbaar waren. Deze fonts worden de 'text companion' fonts genoemd (`tc` fonts). Hierin kunnen zich ook symbolen bevinden die geen plaats hebben in de `ec` fonts.

2 Het gebruik van de ISO-8859-1 set

Wanneer bepaalde characters moeten worden afgedrukt kunnen ze in fonttabellen opgezocht worden. Wanneer ze gevonden zijn, kunnen ze expliciet opgenomen worden door het betreffende font te definiëren. Incidenteel is deze methode heel goed bruikbaar, maar voor een gestandaardiseerde characterset als de ISO-8859-1 set zou een meer algemene benadering gewenst zijn.

Het blijkt dat de genoemde `ec` en `tc` fonts alle characters bevatten om de ISO-8859-1 set volledig te kunnen afdrukken. De volgende vragen blijven dan nog over:

1. Hoe voer je ISO-8859-1 characters in in je computer.
2. Hoe druk je ze via TeX op een zo elegant mogelijke manier af.

2.1 Het invoeren van ISO-8859-1 characters

Deze sectie is noodzakelijkerwijs nogal systeemafhankelijk. Ik werk voornamelijk onder Linux met emacs en vi als editors.

Een deel van de ISO-8859-1 characters, die met accenten, kunnen natuurlijk op de gebruikelijke manier worden opgegeven, é kan nog steeds worden ingevoerd als `\'e`. Dit gaat niet voor alle characters en er moeten dus andere manieren bestaan.

Een standaard toetsenbord kent alleen de mogelijkheid de gewone ASCII characters in te toetsen. Dit is slechts een subset van de ISO standaard. Characters met ASCII code > 127 kunnen niet direct worden opgegeven. Het achterliggende computersysteem zou er dan voor kunnen zorgen dat bepaalde combinaties van tekens worden omgezet in het gewenste character. Hiervoor zijn verscheidene mogelijkheden:

- Onder Linux (en ongetwijfeld onder meer operating systems) kan gebruik gemaakt worden van de `Alt` toets, waarbij meestal eerst gezorgd moet worden dat de `Alt` toets niet wordt afgevangen door een shell voor andere doeleinden. Hiervoor moet onder Linux `C-v`(de `Ctrl` toets tesamen met de `v` toets) gegeven worden.
  De ASCII waarde van de toets die tegelijk wordt ingedrukt met de `Alt` toets wordt met 128 verhoogd. Onder `openwin` kan dit problemen geven. `Alt-q` b.v. gooit (onder Linux) het window weg (ook na `C-v`). Het is mij nog niet gelukt dit uit te zetten. Ik heb niet nog

meer van dit soort combinaties gevonden, maar je weet maar nooit.
Verder is er geen relatie tussen het ASCII character en het character met een waarde 128 groter. B.v. `Alt-"` geeft `\242` in een `xterm`-window. In `vi` wordt dit ¢. De characters hebben duidelijk geen relatie tot elkaar.
In `vi` kan de `Alt` combinatie overigens direct worden ingetoetst, zonder `C-v` ervoor.
Combinaties als `Alt-q` kunnen (in ieder geval onder Linux) niet worden gegeven, en het bijbehorende symbool is dan niet bruikbaar. Bij gebruik van `twm` of `fvwm` in plaats van `openwin` speelt dit probleem niet.

- In `emacs` zijn een aantal bibliotheek functies aanwezig die het gebruik van ISO-8859-1 vereenvoudigen. Deze zijn vanaf emacs 19 beschikbaar gekomen.
  De standaard methode is het geven van `C-q` gevolgd door drie octale cijfers die gelijk zijn aan de charactercode. Het nadeel is dat deze code bekend moet zijn en er geen herkenbare relatie bestaat tussen de code en het character.
  De eerder genoemde methode, waarbij de `Alt` toets wordt gebruikt, is binnen `emacs` ook mogelijk. Deze toets is hier ook in gebruik voor andere functies en moet daarom afgescherm worden. Dit kan door eerst `C-q` te geven. Ook hier kunnen problemen ontstaan doordat characters door het omringende systeem worden afgevangen, zoals `Alt-q`.
  De twee genoemde methoden hebben allebei nadelen. Er is nog een derde invoermethode mogelijk welke gebruik gemaakt van een prefix code n.l.: `C-x8` gevolgd door twee characters die tesamen zo veel mogelijk een indicatie zijn van het nieuwe te vormen character. ¢ ontstaat nu door `C-x8*c` te geven. Deze mogelijkheid moet via een bibliotheek functie worden opgestart. Dit gaat met:
  `Alt-x` load-library ¡enter¿ iso-transl@.
  Standaard wordt na ingave van een niet ASCII character de octale representatie gegeven. Na installeren van de `iso-ascii` bibliotheek functie (`Alt-x` load-library ¡enter¿ iso-ascii@) wordt tussen accolades een indicatie van het teken gegeven. ¢ wordt dan weergegeven als `{c}`.
  Na `Alt-x standard-display-european` wordt het werkelijke teken weergegeven, dit werkt niet altijd, het omringende systeem (b.v. X-Windows) moet dit ondersteunen.

2.2 Alle gegevens bij elkaar

In de tabel aan het einde van dit artikel, worden alle genoemde mogelijkheden op een rij gezet. Zowel het character, als de octale, decimale en hexadecimale representatie worden gegegen. Daarnaast wordt gegeven welke de toets is die tesamen met de `Alt` toets moet worden ingedrukt om het betreffende character te krijgen. Ook de emacs toets combinatie die na `C-x8` gegeven moet worden wordt vermeld. In de laatste kolom wordt de ASCII representatie gegeven die binnen emacs gebruikt kan worden op een scherm waar de echte symbolen niet aanwezig zijn.

Alleen de niet-ASCII characters worden weergegeven, de ASCII characters zelf wijzigen niet wat betreft invoer en weergave.

Wanneer er achter een character `(m)` staat, is het character alleen in mathmode beschikbaar en moet binnen textmode tussen twee `$`'s worden geplaatst worden. Deze characters zijn ook in textmode beschikbaar via de `tc` fonts. Hierop kom ik in de volgende sectie terug.

2.3 Het gebruik binnen LaTeX2e

De wijze waarop de volledige ISO-8859-1 characterset kan worden ingevoerd is in de vorige subsectie duidelijk gemaakt. De vraag die nu nog rest is of TeX daar iets mee kan. Gezien de beloofde tabel zal dat wel vermoed ik.

Bij de bespreking van de `ec` fonts is aangegeven dat het font coderingsschema de naam `T1` heeft. In LaTeX2e is het mogelijk dit op te geven door `T1` op te gegeven als parameter aan de style file `fontenc` d.m.v.
`\usepackage[T1]{fontenc}`.
Hierdoor wordt automatisch van de `dc` fonts gebruik gemaakt. [1] Echter een character positie in de `dc` fonttabel komt i.h.a. niet overeen met haar positie in de ISO-8859-1 characterset. Er moet dus nog een mapping tussen de ISO-set en de `ec` fonttabel gemaakt worden.

Dit kan gebeuren door naast de fontcodering ook de inputcodering op te geven. D.m.v. de style file `inputenc` is dit mogelijk. Deze heeft een parameter `latin1`, synoniem voor ISO-8859-1. Hiermee lijkt het probleem opgelost.

Dit blijkt echter niet het geval te zijn. Er zijn vier characters die wel in de ISO-8859-1 characterset voorkomen maar niet in de `ec` fonts. Ze zijn echter wél in de al eerder genoemde `tc` fonts aanwezig, maar worden niet standaard gebruikt. Na het meegeven van de style file `textcomp` gebeurt dit wél. Deze style file bevat nog veel meer definities welke symbolen die alleen in mathmode beschikbaar zijn in text mode beschikbaar maken. Dit kan o.a. zinvol zijn voor verbatim listings van symbolen. Omdat een symbool óf in math mode óf in textmode beschikbaar is kan dit ongewenste effecten hebben voor de rest van een artikel.

Jörg Knappen de persoon die de `ec` en `tc` fonts onderhoudt heeft speciaal voor verbatim listings van symbolen een speciale versie van `latin1` gemaakt, genaamd `latin1jk`. Wanneer deze in dit artikel zou worden gebruikt kunnen alle symbolen uit de tabel die als mathematisch vermeld staan worden ontdaan van hun dollars omdat ze nu in textmode beschikbaar zijn. Na navraag bij hem bleek dat hij deze speciale versie blijft onderhouden en dat er dus algemeen gebruik van kan worden gemaakt.

De preamble van de LaTeX file welke op dit moment het gewenste resultaat geeft ziet er als volgt uit:

[1] Wanneer de `ec` fonts aanwezig zijn kan standaard voor deze fonts worden gekozen. Zie hiervoor de installatie beschrijving bij de `ec` fonts.

```
\documentclass[11pt]{artikel3}
\usepackage[T1]{fontenc}
\usepackage{textcomp}
\usepackage[latin1]{inputenc}
\usepackage[dutch]{babel}
```

Nu volgt de eerder genoemde tabel met alle gegevens:

Char	Oct	Dec	Hex	C-x 8	Alt-	ISO-ASCII
	240	160	A0	*␣	␣	
¡	241	161	A1	*!	!	{!}
¢	242	162	A2	*c	"	{c}
£	243	163	A3	*L	#	{GBP}
¤	244	164	A4	*$	$	{$}
¥	245	165	A5	*Y	%	{JPY}
¦	246	166	A6	*|	&	{|}
§	247	167	A7	*S	’	{S}
¨	250	168	A8	""	(	{"}
©	251	169	A9	*C	)	{C}
^{a} (m)	252	170	AA	_a	*	{_a}
«	253	171	AB	*<	+	{<<}
¬ (m)	254	172	AC	~~	,	{~}
	255	173	AD	*-	-	{-}
®	256	174	AE	*R	.	{R}
¯	257	175	AF	*=	/	{=}
° (m)	260	176	B0	*o	0	{o}
± (m)	261	177	B1	*+	1	{+-}
^{2} (m)	262	178	B2	^2	2	{2}
^{3} (m)	263	179	B3	^3	3	{3}
´	264	180	B4	’’	4	{’}
μ (m)	265	181	B5	*u	5	{u}
¶	266	182	B6	*P	6	{P}
·	267	183	B7	*.	7	{.}
¸	270	184	B8	,,	8	{,}
^{1} (m)	271	185	B9	^1	9	{1}
^{o} (m)	272	186	BA	_o	:	{_o}
»	273	187	BB	*>	;	}>>}
¼	274	188	BC	1/4	<	{1/4}
½	275	189	BD	1/2	=	{1/2}
¾	276	190	BE	3/4	>	{3/4}
¿	277	191	BF	*?	?	{?}
À	300	192	C0	‘A	@	{‘A}
Á	301	193	C1	’A	A	{’A}
Â	302	194	C2	^A	B	{^A}
Ã	303	195	C3	~A	C	{~A}
Ä	304	196	C4	"A	D	{"A}
Å	305	197	C5	/A	E	{AA}
Æ	306	198	C6	/E	F	{AE}
Ç	307	199	C7	,C	G	{,C}
È	310	200	C8	‘E	H	{‘E}
É	311	201	C9	’E	I	{’E}
Ê	312	202	CA	^E	J	{^E}
Ë	313	203	CB	"E	K	{"E}
Ì	314	204	CC	‘I	L	{‘I}
Í	315	205	CD	’I	M	{’I}
Î	316	206	CE	^I	N	{^I}
Ï	317	207	CF	"I	O	{"I}
	320	208	D0	~D	P	{-D}
Ñ	321	209	D1	~N	Q	{~N}
Ò	322	210	D2	‘O	R	{‘O}

Char	Oct	Dec	Hex	C-x 8	Alt-	ISO-ASCII
Ó	323	211	D3	’O	S	{’O}
Ô	324	212	D4	^O	T	{^O}
Õ	325	213	D5	~O	U	{~O}
Ö	326	214	D6	"O	V	{"O}
× (m)	327	215	D7	*x	W	{x}
Ø	330	216	D8	/O	X	{/O}
Ù	331	217	D9	‘U	Y	{‘U}
Ú	332	218	DA	’U	Z	{’U}
Û	333	219	DB	^U	[	{^U}
Ü	334	220	DC	"U	\	{"U}
Ý	335	221	DD	’Y	]	{’Y}
	336	222	DE	~T	^	{TH}
ß	337	223	DF	"s	_	{ss}
à	340	224	E0	‘a	‘	{‘a}
á	341	225	E1	’a	a	{’a}
â	342	226	E2	^a	b	{^a}
ã	343	227	E3	~a	c	{~a}
ä	344	228	E4	"a	d	{"a}
å	345	229	E5	/a	e	{aa}
æ	346	230	E6	/e	f	{ae}
ç	347	231	E7	,c	g	{,c}
è	350	232	E8	‘e	h	{‘e}
é	351	233	E9	’e	i	{’e}
ê	352	234	EA	^e	j	{^e}
ë	353	235	EB	"e	k	{"e}
ì	354	236	EC	‘i	l	{‘i}
í	355	237	ED	’i	m	{’i}
î	356	238	EE	^i	n	{^i}
ï	357	239	EF	"i	o	{"i}
	360	240	F0	~d	p	{-d}
ñ	361	241	F1	~n	q	{~n}
ò	362	242	F2	‘o	r	{‘o}
ó	363	243	F3	’o	s	{’o}
ô	364	244	F4	^o	t	{^o}
õ	365	245	F5	~o	u	{~o}
ö	366	246	F6	"o	v	{"o}
÷ (m)	367	247	F7	//	w	{/}
ø	370	248	F8	/o	x	{/o}
ù	371	249	F9	‘u	y	{‘u}
ú	372	250	FA	’u	z	{’u}
û	373	251	FB	^u	{	{^u}
ü	374	252	FC	"u	|	^"u}
ý	375	253	FD	’y	}	{’y}
	376	254	FE	~t	~	{th}
ÿ	377	255	FF	"y	DEL	{"y}