Column

Unicode ook voor blinden?

Henk Gianotten

Het logo van Unicode.

Unicode is de coderingsstandaard achter technieken als HTML, XML email en tekstverwerking voor documenten, tijdschriften en boeken. Oude gedigitaliseerde boeken en nieuwe teksten kunnen zo zichtbaar gemaakt worden op papier, beeldscherm en mobiel en door een knappe aanvullende techniek geschikt worden gemaakt voor dyslectici, visueel gehandicapten én laaggeletterden.

Loodzetmachines werden in hun nadagen gevoed met zogenaamde Teletype-ponsbanden, waardoor een hogere productie gerealiseerd kon worden. De digitale informatie over de lettertekens en spaties stond op een ponsband en de codering van die band werd door de computer en zetmachine gelezen en verwerkt. De codering in een zeskanaals papieren ponsband was gebaseerd op het toetsenbord van de regelzetmachine en was afhankelijk van zowel zetmachinetype, taal als toepassingsgebied.

Bij de latere fotozetsystemen was het niet veel anders, want de coderingen op de digitale dragers als cassette, floppy of cd-rom waren slechts gedeeltelijk gestandaardiseerd. Zetterijen moesten daarom dataconversie plegen en de digitale input van uitgevers en auteurs 'kraken' om de juiste coderingen voor de fotozetmachines te genereren. Pas met PostScript en Desk Top Publishing in 1985 ging het iets beter dankzij standaardisering van de beschrijvingstaal, maar ook toen waren de datadragers van Windows en Apple computers verschillend en gebruikte men verschillende font-indelingen voor beide systemen.

In die tijd ontwierp het Amerikaanse Xerox een veelbelovend nieuw computersysteem met beeldschermweergave en begonnen het leger en de universiteiten in de Verenigde Staten onderling emailberichten uit te wisselen. Dat gebeurde met de toenmalige standaard ASCII+, die slechts 256 letter- en stuurcodes had. In ons land gebruikten we meestal ASCII met de uitbreiding Latin-1, waardoor we meer geaccentueerde letters konden zetten. Xerox werkte samen met Apple en beide bedrijven zagen in dat een internationale uitwisseling van meerdere talen nodig was voor wereldwijd verkoopbare toepassingen. Ook wetenschappers wilden hun documenten – inclusief speciale tekens in allerlei talen en scriptsystemen – betrouwbaar kunnen weergeven. Unicode, onder welke naam het coderingssysteem in 1991 werd geïntroduceerd, voorzag in die behoefte. In de loop der jaren is het de wereldwijde standaard geworden voor het coderen van teksten, die vrijwel alle letters, tekens en symbolen kan omvatten.

De drie coderingsvormen van Unicode: UTF-8, UTF-16 en UTF-32. In Europa en Amerika hanteren we vrijwel altijd UTF-8, waarop ook HTML 4 en 5 en email zijn gebaseerd.

Bedrijven als Adobe, Apple, IBM en Microsoft zegden hun medewerking toe en ook de ISO-organisatie en veel onderzoeksinstituten en universiteiten verwelkomden dit ambitieuze initiatief om alle bestaande coderingstechnieken voor alle relevante talen te vervangen door één overkoepelende, die theoretisch in staat is meer dan een miljoen verschillende tekens weer te geven. Dat zijn naast letters, cijfers en leestekens bijvoorbeeld ook chemische en mathematische tekens, schaak-, bridge- en damtekens, muzieknoten en braille-tekens.

Het heeft bijna tien jaar geduurd voordat de IT-industrie overtuigd was van de voordelen. Ook de leveranciers van digitale letters hadden die tijd nodig om naar Unicode fonts te kunnen omschakelen. De oude fonts moesten vervangen worden door de nieuwe platformonafhankelijke mogelijkheden van OpenType fonts en het vervangen en converteren van miljoenen bestanden en fonts was een extreem complexe en tijdrovende zaak. En ook alle software bij uitgevers, auteurs, ontwerpers en drukkers moest gewijzigd worden.

Internet dwong ons

Vanaf de opkomst van internet, websites en email was het duidelijk dat de oude, veelgebruikte encodings van Apple, IBM en Windows niet meer voldeden en vervangen moesten worden. Het gevolg was een veelheid van oude en nieuwe coderingen die soms tegelijkertijd gebruikt moesten worden in één systeem. Dat was onvermijdelijk, omdat alle leveranciers van hard- en software én de gebruikers de oude email- en tekstopslagsystemen in de lucht moesten houden, terwijl men de nieuwe componenten moest invoeren én instrueren.

Unicode is nu dé wereldwijde standaard geworden met drie coderingsvormen die als UTF-8, UTF-16 en UTF-32 gedefinieerd zijn. Voor de westerse wereld is UTF-8 (de acht bit-versie) de standaard geworden voor o.a. email, HTML-4 en 5, XML, Java, OS, webfonts en

Schröder
Schroeder
Schr%der

021D0 E28790 ⇐	021D1 E28791 ⇑	021D2 E28792 ⇒	021D3 E28793 ⇓
⇐	⇑	⇒	⇓
021E0 E287A0 ⇠	021E1 E287A1 ⇡	021E2 E287A2 ⇢	021E3 E287A3 ⇣
⇠	⇡	⇢	⇣

Een groepje pijlen uit de set voor speciale tekens.

Model	2018	2019	201A	201B	201C	201D	201E	201F
Rotated model (curly glyph style)	‘	’	‚	‛	“	”	„	‟
Rotated model (wedge glyph style)	❛	❜	❜	❛	❝	❞	❞	❞
Mirrored model (Tahoma, Verdana)	‘	’	‚	‘	“	”	„	“

Verschillende soorten aanhalingstekens.

browsers. Banken en andere dienstverleners moesten overschakelen, maar dat gebeurt traag – lang niet bij alle banken kun je voor digitale overschrijvingen voldoende relevante geaccentueerde letters gebruiken. Sinds 2003 is Unicode 4.0 in gebruik en waren ruim 96.000 tekens beschikbaar. Versie 6.0 kende ruim 109.000 tekens en de huidige versie 11.0 ruim 137.000 stuks. Daar vallen Chinees, Japans, Koreaans en andere schriften uit het Verre Oosten onder, evenals een grote variëteit aan emoji-tekens die op smartphones worden gebruikt om gevoelens digitaal over te brengen.

Tekens met een trema werden vroeger vaak fout weergegeven in emails. Soms veranderde men daarom -ö in -oe.

onder
Alle zeventien in gebruik zijnde relatieve spaties zijn als een blok opgenomen. Een opmaakprogramma kan ook absolute spaties toevoegen.

rechtsonder
Het lijvige handboek *Unicode Standard 4.0* toont op 1464 pagina's ruim 96.000 tekens met hun toelichtingen. Recente versies verschijnen niet meer op papier, maar worden op de websites zichtbaar gemaakt. Ook de omvangrijke PDF's zijn daar te downloaden.

Wereldwijde normen voor toegankelijkheid

De VN stelde begin deze eeuw de eis dat ook gehandicapten toegang moesten krijgen tot digitale informatie. Op haar beurt besloot de EU jaren geleden dat voor documenten, websites en smartphones van overheidsinstellingen de teksten en documenten moeten voldoen aan de WCAG-normen voor toegankelijkheid of drempelvrijheid. Ook Nederland heeft zich verplicht om vanaf 23 september 2019 de drempelvrijheid voor digitale informatie te garanderen. Met zogenaamde AT (Assistive Technologies) en TTS (Tekst To Speech) kunnen blinden en andere visueel gehandicapten teksten laten voorlezen. Ook analfabeten en niet-Nederlands sprekende ingezetenen kunnen met behulp van relatief eenvoudige hulpmiddelen teksten laten voorlezen. Men heeft daarbij de keuze uit bv taal, toonhoogte, tempo en zelfs stemsoort.

Het voorlezen gaat met behulp van de in het OS ingebouwde TTS-mogelijkheden die Android, Apple en Microsoft al bieden. De gestructureerde documenten in PDF-formaat moeten ook aan bepaalde technische eisen voldoen, zoals het insluiten van OT-fonts of TrueType fonts die aan UTF-8 voldoen. Alle teksten en dus ook die in de PDF moeten voldoen aan die coderingen van UTF-8; dat laatste is vooral nodig omdat de techniek relatief goedkoop moet blijven en men dus additionele coderingsvarianten uitsluit. Voor overheden, semi-overheden en bedrijfsleven (denk aan websites die ook aan de toegankelijkheidseisen WCAG moeten voldoen) wordt het nog een hele klus om tijdig bestanden te converteren en te controleren op de toepassing van deze ISO-coderingsstandaard. Uiteindelijk hoopt de overheid dat bijvoorbeeld blinden voldoende toegang krijgen tot documentatie. Een ander belangrijk doel is om laaggeletterden en buitenlanders met behulp van deze technieken op termijn toegang te geven tot informatie van artsen, ziekenhuizen, sociale diensten en gemeentehuizen. Of gewoon thuis, waar het document geconsumeerd kan worden als een podcast. Of nog mooier, een variant van het e-book met meertalige documentatie die zowel gelezen als beluisterd kan worden. De infrastructuur daarvoor komt beschikbaar en de standaards zijn aanwezig. De uitvinders van PDF konden dat 25 jaar geleden hoogstwaarschijnlijk niet bevroeden.

Zie ook www.unicode.org en www.decodeunicode.org

Code	Name
U+0020	SPACE
U+00A0	NO-BREAK SPACE
U+1680	OGHAM SPACE MARK
U+2000	EN QUAD
U+2001	EM QUAD
U+2002	EN SPACE
U+2003	EM SPACE
U+2004	THREE-PER-EM SPACE
U+2005	FOUR-PER-EM SPACE
U+2006	SIX-PER-EM SPACE
U+2007	FIGURE SPACE
U+2008	PUNCTUATION SPACE
U+2009	THIN SPACE
U+200A	HAIR SPACE
U+202F	NARROW NO-BREAK SPACE
U+205F	MEDIUM MATHEMATICAL SPACE
U+3000	IDEOGRAPHIC SPACE