Wat is UTF-8 en waarom is het belangrijk?

Henk Vanstappen, PACKED vzw

Iedereen kent het wel: een e-mail, website of database waarin plots vreemde tekens opduiken: François heet plots FranÃ§ois. Bij nader toezien blijkt het telkens te gaan om dezelfde tekens die door het programma fout worden begrepen. Hoe komt dat en hoe vermijd je het?

Computers begrijpen alleen enen en nullen. Letters en andere tekens worden opgeslagen door ze als een binair getal te coderen: een a wordt volgens het ASCII-systeem bijvoorbeeld gecodeerd als 1100001. Er zijn dus zeven enen of nullen nodig om de letter a te coderen. Maar zeven bits volstaan niet om alle tekens uit alle talen te coderen. Als oplossing werden per taal verschillende coderingssystemen vastgelegd. Dat leidt echter tot verwarring: 11000110 stelt in het ene systeem de ligatuur Æ voor, maar in een ander de Slavische Ć. Vandaar dus dat sommige teksten er wat raar uit zien.

UNICODE

Om dit te verhelpen werd Unicode ontwikkeld, wat je je kan voorstellen als één gigantische tabel met alle mogelijk tekens uit alle mogelijke talen. Voor die verzameling tekens zijn enkele universele coderingssystemen vastgelegd. De belangrijkste is UTF-8, dat acht tot tweeëndertig bits gebruikt om een teken op te slaan. Voor de meest voorkomende tekens worden slechts acht bits gebruikt, waardoor UTF-8 toch bijna even compact is als ASCII. Alle gewone ASCII-codes maken deel uit van de UTF-8 character set.

UTF-8: eenduidig coderen van tekst is nooit zo gemakkelijk geweest.
Foto: The U.S. National Archives.

UTF-8 is uitgegroeid tot het standaard coderingssysteem voor het uitwisselen van digitale tekst, zoals e-mails, webpagina's of databasegegevens. Het biedt je de beste garantie op een foutloze weergave van tekst. De meeste teksteditors ondersteunen UTF-8 en geven je bij het opslaan de keuze uit een aantal coderingssystemen. Wanneer je voor UTF-8 kiest, zijn er weinig problemen te verwachten.

> "UTF-8 IS UITGEGROEID TOT HET STANDAARD CODERINGSSYSTEEM VOOR HET UITWISSELEN VAN DIGITALE TEKST, ZOALS E-MAILS, WEBPAGINA'S OF DATABASEGEGEVENS."

UTF-8 VALIDEREN

Hoe weet je volgens welk systeem een tekst gecodeeerd is? In het begin van een XML- of een HTML-bestand wordt de codering aangegeven, maar dat is niet altijd betrouwbaar. Sommige programma's zijn in staat om de tekstcodering te 'raden'. Voorbeelden zijn Notepad++ of TextWrangler die de codering vermelden en toelaten een bestand in een andere codering te openen of te bewaren. Een ander hulpmiddel is de UTF-8 module van JHOVEview. Als het werkelijk om een codering in UTF-8 gaat, valideert dit programma het bestand als *Well-formed and valid.* Maar zeker ben je nooit: voor een computer is FranÃ§ois een geldige tekenreeks, en mogelijk heeft de auteur immers bewust rare tekens gebruikt.

> **Dit artikel werd bezorgd door PACKED vzw dat als expertisecentrum digitaal erfgoed kennis, ervaring en deskundigheid omtrent digitaal erfgoed centraliseert en verspreidt. Vragen voor PACKED vzw zijn welkom via info@packed.be.**