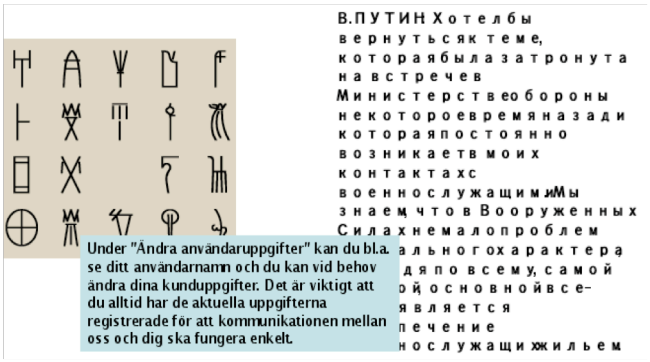
ï ə ɲ

Domeinnaam Debat 2006, een column

30 November 2005
Olaf M. Kolkman
olaf@NLnetLabs.nl

Ik ben gevraagd om de discussie over IDN met een column in te leiden. Ik noem mijzelf soms DNS protocolspecialist en een technische presentatie over Unicode, Nampeprep, stringprep en normalisatie zou dus voor de hand liggen. Dit wordt niet zo'n presentatie. Ik neem als uitgangspunt de aanbeveling van SIDN en stel hardop wat vragen. Ik chargeer zo nu en dan een beetje, het blijft tenslotte een column.

Veel protocollen op het Internet gaan uit van 'hostnames'. Hostnames hebben de restrictie dat ze worden uitgedrukt door 26 Latijnse letters, 10 cijfers, en een streepje. De namen van webservers zijn 'hostnames', de namen aan de rechterzijde van het apenstaartje in een e-mail adres zijn 'hostnames'. Hostnames zijn een deel verzameling der domeinnamen, maar dat is een technisch detail waar ik niet verder op in ga. Wordt niet verward doordat de termen domein en hostname door elkaar gebruikt worden.



Het International Domain Names, of IDN, protocol is uitgevonden om 'hostnames' weer te geven in een schrift dat afwijkt van 26 Latijnse letters, 10 cijfers en een streepje. Er zijn nogal wat schriften die afwijken van die beperkte karakterset; Japans, Chinees, Linear B, Zweeds, Russisch, &c, &c. IDN maakt het mogelijk om domeinnamen op in verschillende schriften te representeren en voorziet daarmee in een economisch behoefte. Niet voor de gebruikers van Linear-B maar voor Japans en Mandarijn des te meer.

قصيدة.شركة.nu

↕

xn--ogblr5czb.xn--ogbpi5d.nu

IDN is een afspraak over hoe ingetypte tekst in een 'vreemd' schrift wordt omgezet naar tekens die mogen voorkomen in een 'hostname'. Het definieert ook hoe de 'hostname' weer naar de oorspronkelijke set glyphen wordt terug getransformeerd. IDN wordt geïmplementeerd op applicatie niveau. Voor gebruikers van een applicatie zonder IDN ziet de domeinnaam er ook uit als een klassieke domeinnaam die de combinatie "xn--" en wat onbegrijpelijke letter combinaties bevat.

Men zou alle denkbare karakters kunnen toelaten. Dat is niet echt praktisch, het lijkt dus redelijk om keuzes te maken met betrekking tot het gebruik van de karaktersets. Die keuzes zijn commercieel, economisch en cultureel.

Ten eerste zullen we moeten kiezen welke schriften we gebruiken.

Den Nederlandsche Internet Gebruiker

Bron: UNOX reclame

In haar aanbeveling stelt SIDN dat het zijn afbakening definieert ten behoeve van de Nederlandse Internet gemeenschap. Ik weet niet waar die Nederlandse Internet gemeenschap uit bestaat. Zijn dat de mensen die surfen naar Nederlandse domeinnamen, de mensen die een account hebben bij een Nederlandse provider, de mensen die in het Nederlands communiceren, de mensen die in Nederland zaken doen, de mensen met een Nederlands paspoort die zich op het Internet begeven? U begrijpt dat iedereen die zich aan een dergelijke definitie waagt een goed politicus moet zijn.

Uitgaande van de Nederlandse taal wordt er gekozen voor het 'Latin' schrift. Dat schrift bevat inderdaad alle karakters die we het Nederlands, het Frysk, het Haags en het Limburgs gebruiken. Het omvat zelfs karakters die in het Turks gebruikelijk zijn. Maar waar moet nou de islamitische slagerij om de hoek terecht? Kan hij zijn bedrijfsnaam registreren in het schrift wat zijn klanten appelleert: Arabisch. Mijn islamitische slagerij is ingeschreven in de kamer van koophandel, heeft een Nederlandse Internet provider, slacht Nederlandse geiten en is bovendien oud-Hollandsch goedkoop. Is hij geen deel van de Nederlandse Internet gemeenschap? Al is een keuze voor 'Latin' is niet geheel onlogisch, er wordt wel een culturele grens getrokken die misschien opgerekt moet worden. Wie neemt de verantwoording voor die keuze?

Als er dan al voor een bepaald schrift is gekozen, zijn we dan klaar voor de komende 10 jaar? Kan een toekomstige Europese richtlijn over het gebruik van schriften binnen de lidstaten worden uitgesloten? Het lijkt me geen slecht idee om als randvoorwaarde aan de invoering van IDN te stellen dat nieuwe schriften later kunnen worden ingevoerd.

Welke schriften, één of meer, we ook kiezen we kunnen een aantal karakters uitsluiten dankzij technische limitaties van het IDN protocol: geen spaties, samengestelde karakters, &c. Daarnaast worden de hoofdletters die de door IDN gedefinieerde transformatie niet overleven uitgesloten. Het ruimt lekker op en er is wat mij betreft weinig willekeur in deze keuze.

Niet technische keuzes komen weer om de hoek kijken als we de visueel verwarrende karakters uitsluiten. Die karakters worden in het Engels homoglyphen genoemd. Een Nederlandse variant van dit Latinisme heb ik niet kunnen achterhalen. Maar dat terzijde.
ICANN beveelt aan dat verschillende schriften niet door elkaar gebruikt mogen worden.
Dat de uitbaters van undutchables.nl daardoor hun merknaam niet als UИDUTÇABLΣS.nl onder kunnen registreren is jammer, maar we kunnen de schuld bij de ICANN aanbeveling leggen. Er moet wel tussen verschillende varianten van het Latin schrift worden gekozen, dat is een detail.

Blijft de angst dat er homoglyphen binnen een karakterset voorkomen. "paypal" en "pąypal" (geschreven met een 'a' met 'ogonek') lijken veel op elkaar. Dat verschil kan door booswichten wordt uitgebuit. Op basis van mogelijke verwarring en het niet aanwezig zijn van karakters in het Nederlands wordt in de aanbeveling van SIDN de hele "Extended Latin A" karakterset weggestreept. Sluiten we daardoor de Nederlandse Internet gebruiker uit? Ik weet het niet, ik weet immers niet wie die gebruiker is. Is de restrictie terecht, ook dat weet ik niet.

Met het uitsluiten van de 'Extended Latin A' karakters is de angst voor misbruik van homoglyphen niet uit de lucht. Daarom worden in de SIDN aanbeveling, mijns inziens volstrekt willekeurig, nog een aantal karakters weggegooid. Daaronder de ø, de o met de schuine streep. Dat maakt registratie van bløf.nl onmogelijk. Ik vraag me af of de leden van Bløf zich lid van de Nederlandse Internet gemeenschap voelen of niet. Ze zingen in ieder geval wel bloedmooie Nederlandse liedjes. Kunnen de leden van het 'anhangerschap Normaal' nu niet de domeinnaam 'høken.nl' registreren?

Ook de ï, een i met een trema, wordt weggemieterd. Dat is jammer want 'goed-geïmplementeerd.nl' zou ik zelf een interessante domeinnaam vinden. En waarom wordt een i met een trema wel gezien als homoglyph en een a met trema niet? Voor een nietsvermoedende surfer ziet "äbnamro.nl" er waarschijnlijk hetzelfde uit als "abnamro.nl". Als de gebruiker al naar de adresbalk van zijn browser kijkt. En waar blijft de "ç", de c met cedille? Waar moeten we heen met onze in het Turks geschreven namen (de schoolklas van mijn zoon zit er vol mee).

Over de adresbalk gesproken. IDN heeft alleen betrekking over wat er in een adresveld van een applicatie wordt ingetypt. Niet op de tekst in de browser, of in een e-mail.

Wie heeft er recent nog een domeinnaam in getypt in zijn adresbalk? Wie heeft er in één keer foutloos www.domeinnaamdebat2006.nl in zijn browser getypt en wie heeft er voor "domein debat .nl 2006" ge-googled?

Daarnaast vraag ik me af hoeveel mensen er weten hoe je een é, û of een ï moet typen? Heeft invoering van IDN wel nut als je bijna zeker weet dat een groot deel van de Nederlanders Café zonder accent aigu schrijft. Overleeft www.tête-à-tête.nl een radio commercial? Bovendien heeft die adres balk uiteindelijk weinig te maken met wat er in de browser wordt gepresenteerd.

Mijn islamitische slagerij heeft voor zijn Arabisch schrijvende klanten een link staan vanaf een portal voor Arabisch schrijvende Nederlanders. Ik kan die link gewoon volgen zonder dat ik een letter Arabisch in type. Een uurtje willekeurig rondklikken op sites die Japans schrift gebruiken kan ook erg opwindend zijn. Ik kan mijn IJslandse collega Guðmundsson gewoon correct adresseren door te clicken op een link in mijn adresboek. Mijn punt is dat ik een bepaald schrift niet in de adres balk hoeft te kunnen typen om in je eigen schrift te kunnen mailen en surfen. Die adresbalk is zo belangrijk niet.

Maar laten we dit nu eens allemaal terzijde schuiven en eens kijken naar de economische zijde van het verhaal.

Hoeveel gaat de invoering van IDN kosten? Een paar zaken die in ieder geval geld gaan kosten is de intergratie in de Whois Database,in de Whois clients, in het Billing systeem, in de systemen waarmede de leden met SIDN communiceren en in de systemen waarmee de leden zaken doen met hun klanten. Al deze systemen moeten zo gebouwd worden dat andere karakters op termijn moeten worden kunnen ingevoerd. Met andere woorden op bijna alle systemen moet IDN worden geïmplementeerd.

Juridische kosten gaan omhoog, ook al probeer je homoglyphen uit te sluiten je gaat te maken krijgen met homografen. Zijn "cafe-bluf" en "café-bløf" verschillende merken, betekenen ze hetzelfde? De invoering van meer variatie in het schrift zal leiden tot meer variatie in interpretatie en dus in hogere juridische rekeningen.

Stel we gaan al die kosten naar rato omslaan naar die enkele web site houder die IDN gaat gebruiken. Is de prijs van een domeinnaam dan nog wel reëel? Hoeveel betaald u voor "reëel.nl"? Gaan we er met zijn allen voor betalen? Is er een economische behoefte of gaat het alleen om het weergeven van beeldmerken in het adresveld van e-mail programma's en browsers?

U begrijpt het van mij hoeft het niet zo nodig. Hebben we ons overigens al afgevraagd welk probleem we proberen op te lossen? Waarom we het in de eerste plaats over IDN hebben?