

Internationalized **D**omain **N**ames

I **D** **N** **s**

Plan:

🌐 IDN, IDNA, Unicode

🌐 Stringprep, Nameprep i Punycode.

🌐 Operacje ToUnicode, ToASCII.

🌐 Wprowadzenie IDNA.

IDN

czyli Internationalized Domain Name; jest to nazwa domenowa, którą możemy zapisać używając znaków zaczerpniętych z repertuaru znaków Unikodu.

IDNA

czyli Internationalizing Domain Names in Application; jest to mechanizm odpowiadający za utrzymanie i używanie IDN w „standardowym stylu” (RFC 3490).

Unikod

(*ang. Unicode*) przypisuje każdemu znakowi unikalny numer (kod numeryczny, *ang. code point*), niezależny od używanej platformy, programu czy języka. Unikod jest nowoczesnym sposobem kodowania obejmującym znaki używane na całym świecie (np. polskie ogonki, hieroglify, cyrylicę), symbole muzyczne, techniczne, fonetyczne i inne często spotykane. Ważną cechą Unikodu jest fakt, że pierwsze 128 znaków odpowiada kodom ASCII (zakres 00..7F).

Cechy IDNA

- 🌐 Pozwala na używanie w nazwach domen internetowych znaków „non-ASCII” reprezentowanych za pomocą znaków ASCII.
- 🌐 IDNA nie wprowadza żadnych zmian do infrastruktury Internetu. Żaden z istniejących protokołów nie musi być modyfikowany aby móc używać IDN.
- 🌐 Protokoły niższych warstw nie muszą być „świadome” użycia IDN w aplikacji.
- 🌐 IDNA ma zastosowanie tylko do aplikacji użytkownika (IDNA ingeruje tylko w warstwę aplikacji modelu ISO OSI), np.: przeglądarki internetowe, klienci poczty elektronicznej, klienci FTP itp.

Możliwe problemy i komplikacje

- 🌐 Użytkownik IDNA musi wiedzieć jak dokładnie wprowadzić daną domenę (literka po literce).
- 🌐 Wprowadzenie dużego repertuaru znaków może powodować dużą ilość błędów (literówek) w pisaniu nazw domen.
- 🌐 Możliwość błędnego odczytania domeny ze względu na podobieństwo niektórych znaków.
- 🌐 Problemy we wprowadzeniu poprawnej nazwy domeny w oparciu o wizualną albo oralną informację (np. billboard i radio).
- 🌐 Może powstać wiele podobnie wyglądających lub brzmiących nazw.

Co „piszczy” w środku?

🌐 IDNA zawiera się tylko w aplikacji.

🌐 Wykonywane są dwie operacje:

🌐 ToASCII – jest wykonywana przed wysłaniem IDN do czegoś, co oczekuje nazw w kodzie ASCII (np. resolver) lub zapisuje IDN w miejsce, gdzie oczekiwana jest nazwa w kodzie ASCII (np. plik konfiguracyjny serwera DNS).

🌐 ToUnicode – jest wykonywana przed wyświetleniem IDN użytkownikowi.

🌐 IDNA wymaga zaimplementowania Nameprep’a, który jest profilem Stringprep’a, i Punycode’a.

Stringprep

Stringprep (RFC 3454) - ogólny algorytm przygotowania tekstu Unikodowego przed zamianą go na ciąg znaków ASCII (tzn. algorytm, który może być profilowany - dostosowywany do różnych potrzeb). Draft określa kilkanaście tabel (mapowania, znaki zakazane, tablice "dwukierunkowe"), które mogą być zastosowane w algorytmach opartych na stringprep'ie (m.in. nameprep).

Nameprep

Nameprep (RFC 3491) jest profilem stringprep'a (w drafcie określono dokładnie, które z tabel mapowania, znaków zakazanych etc., są zastosowane). Reguły przetwarzania zostały wybrane na potrzeby kodowania nazw domenowych, a nie zwykłego tekstu (=> dozwolone znaki zgodne z RFC1035 [a-zA-Z0-9-]).

Punycode

Punycode (RFC 3492) to instancja algorytmu Bootstring (ma inne wartości parametrów inicjalizujących). Algorytm pozwala jednoznacznie odwzorować ciągi znaków stworzone z większego zbioru znaków w ciągi złożone z mniejszego zbioru znaków.

Cechy Punycode'a

- 🌐 **Kompletność** – każdy ciąg znaków można zaprezentować przy pomocy zbioru znaków podstawowych.
- 🌐 **Unikalność** - istnieje co najwyżej jedno odwzorowanie ciągu rozszerzonego w podstawowy.
- 🌐 **Odwracalność**.
- 🌐 **Efektywność kodowania** - małe wydłużenie ciągu znaków po zakodowaniu (stosunek długość podstawowego ciągu do długości ciągu rozszerzonego po kodowaniu jest niewielki). Jest to bardzo ważna cecha , ponieważ RFC1034 ogranicza długość etykiety domenowej do 63 znaków.
- 🌐 **Prostota algorytmu** - łatwy do zaimplementowania.
- 🌐 **Czytelność** - znaki ze zbioru podstawowego pojawiające się w ciągu rozszerzonym są reprezentowane przez nie same.

Wymagania i zastosowanie

🌐 Wszędzie tam, gdzie kropki występują jako separatory etykiet następujące znaki: U+002E (full stop), U+3002 (ideographic full stop), U+FF0E (fullwidth full stop), U+FF61 (halfwidth ideographic full stop) muszą być rozpoznawane jako kropki.

🌐 Tam gdzie jest wymagana domena w kodzie ASCII musi być podana domena w kodzie ASCII.

🌐 Etykiety ACE (*ang. ACE labels* → xn--zaglja-cxa0mpa5p6q5a80a6ota.com.pl) muszą być ukryte przed użytkownikiem, za każdym razem, kiedy to tylko możliwe; chyba że forma etykiety domeny ACE jest wyraźnie żądana.

www.xn--zaglja-cxa0mpa5p6q5a80a6ota.com.pl → www.zażółćgęślajaźń.com.pl

www.xn--bzdzigwawddystydzie-fgc15k68aba.pl → www.bzdziągawdźdżystydzień.pl

🌐 Kiedykolwiek etykiety są porównywane, trzeba sprawdzić czy są sobie równoważne, czyli należy porównywać ich formę ASCII (ACE labels).

Jak działa IDNA?

1. Podajemy nazwę domeny. Jeżeli domena jest przedstawiona za pomocą innego zestawu znaków niż Unikod lub ASCII, to najpierw jest przekodowywana do Unikodu.
2. Rozróżnianie czy jest to „stored string” czy „query string”.
3. Podana domena jest dzielona na osobne etykiety (etykiety nie zawierają separatorów, czyli kropek).
4. Etykieta sprawdzana jest pod względem spełniania wymogów dotyczących nazw hostów i domen (zgodność z standardami STD3 i STD13).
5. Przetwarzanie kolejno etykiet i wybór operacji jaka ma być użyta, ToUnicode czy ToASCII.
6. Jeżeli została wybrana operacja ToASCII, to wszystkie kropki użyte jako separatory etykiet zamienione są na U+002E (full stop).

Operacja ToASCII

🌐 Operacja ToASCII pobiera sekwencje kodów numerycznych (*ang. code points*) Unikodu, które tworzą jedną etykietę i przekształca je w sekwencje kodów z zakresu ASCII (0..7F).

🌐 Jeżeli operacja się powiedzie, to oryginalna sekwencja kodów numerycznych i ta otrzymana po konwersji są równoważnymi etykietami.

www.zaźółćgęślajażń.com.pl = www.xn--zaglja-cxa0mpa5p6q5a80a6ota.com.pl

🌐 Jeżeli operacja ToASCII zawiedzie, to znaczy że domena nie może być użyta jako domena IDN.

🌐 Operacja ToASCII nigdy nie zmienia kodów, które są w zakresie ASCII.

Jak działa operacja ToASCII?

1. Wprowadzamy domenę.
2. Jeżeli została wprowadzona w lokalnym zestawie znaków np. ISO 8859-2, to przekodowywana jest do Unikodu.
3. Zamiana znaków na ich kody numeryczne.
4. Sprawdzanie czy dana sekwencja kodów numerycznych jest poprawna, tzn. czy nie zawiera kodów znaków zabronionych (zgodność z STD3 i STD13). Tutaj zostaje użyty Nameprep.
5. Jeżeli domena przeszła test, to przekodowywana jest za pomocą algorytmu Punycode.
6. Do przekodowanej etykiety dodawany jest prefiks ACE „xn--”.
7. Sprawdzanie, czy liczba kodów numerycznych nie jest większa niż 63.

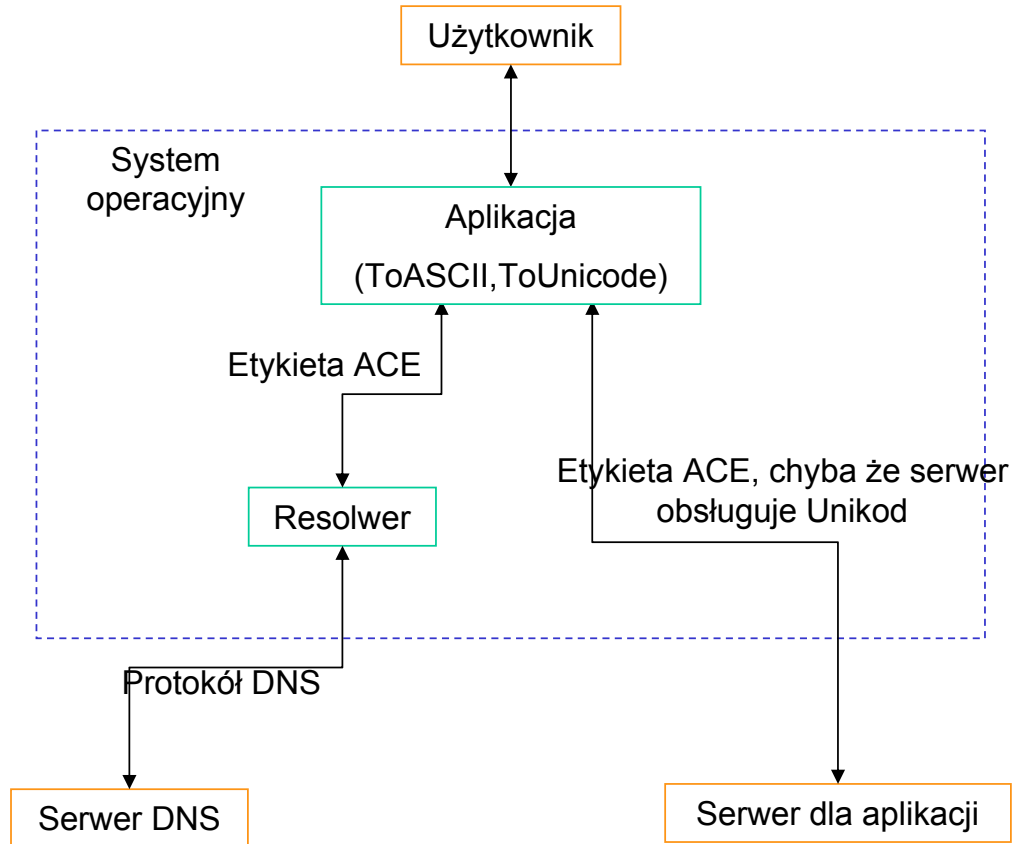
Operacja ToUnicode

- 🌐 Operacja ToUnicode pobiera sekwencje kodów ASCII, które tworzą etykietę i zwraca sekwencje kodów Unikodu.
- 🌐 Jeżeli na wejściu do tej operacji jest etykieta z prefiksem ACE (xn--), wtedy wynik operacji stanowi równoważna etykieta IDN. Jeżeli natomiast nie, to na wyjściu jest nie zmieniona etykieta.
- 🌐 Operacja ToUnicode nigdy nie zawodzi.

Jak działa operacja ToUnicode?

1. Na wejściu operacji podawana jest domena w kodzie ASCII.
2. Sprawdzanie czy domena posiada prefiks ACE „xn--”, zapamiętanie sekwencji kodów.
3. Usunięcie prefiksu ACE.
4. Przekodowanie sekwencji kodów za pomocą algorytmu Punycode; zapamiętanie wyniku.
5. Zastosowanie operacji ToASCII.
6. Porównanie wyników z etapu 5 i 2 (sprawdzenie poprawności przekodowania domeny)
7. Wyświetlenie wyniku z etapu 4.

Schemat IDNA



Zasady działania aplikacji wspierającej IDNA

- 🌐 Wszystkie aplikacje pobierające nazwy domen lub hostów z zasobów DNS powinny zostać zmodyfikowane tak, aby móc ukryć etykietę ACE przed użytkownikiem.
- 🌐 Jeżeli system nie potrafi poprawnie wyświetlić domeny po operacji ToUnicode, wówczas powinna zostać wyświetlona jej wersja ACE.
- 🌐 Jeżeli istnieje domena zaczynająca się prefiksem ACE a nie jest IDN, wówczas operacja ToUnicode nie zmienia tej domeny.
- 🌐 Wszystkie aplikacje muszą stosować jednolitą metodę konwersji IDN.

Wprowadzenie IDNA

1. Wprowadzenie IDNA wymusza aktualizację programów, które są w interakcji z domenami np.: przeglądarki internetowe, klienci poczty elektronicznej, klienci ftp, edytory html, edytory tekstu itp.
2. Rozpowszechnienie IDNA zależy głównie od propagacji oprogramowania, które je wspiera.

Koniec
