

NAUKOWA I AKADEMICKA SIEĆ KOMPUTEROWA

Internationalized
Domain
Names

Seminarium dla CBR

Krzysztof Olesik
Andrzej Bartosiewicz

Krzysztof.Olesik@nask.pl
Andrzej.Bartosiewicz@ties.itu.int

Wprowadzenie

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 domen 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).

Implementacja domen IDN opiera się na wersji Unikodu 3.2.

www.rozyczka.pl

Dozwolone znaki:

- myślnik „-”
 - litery a-z
 - cyfry 0-9
-

www.różyczka.pl

Forma unikodowa domeny IDN

prefiks ACE

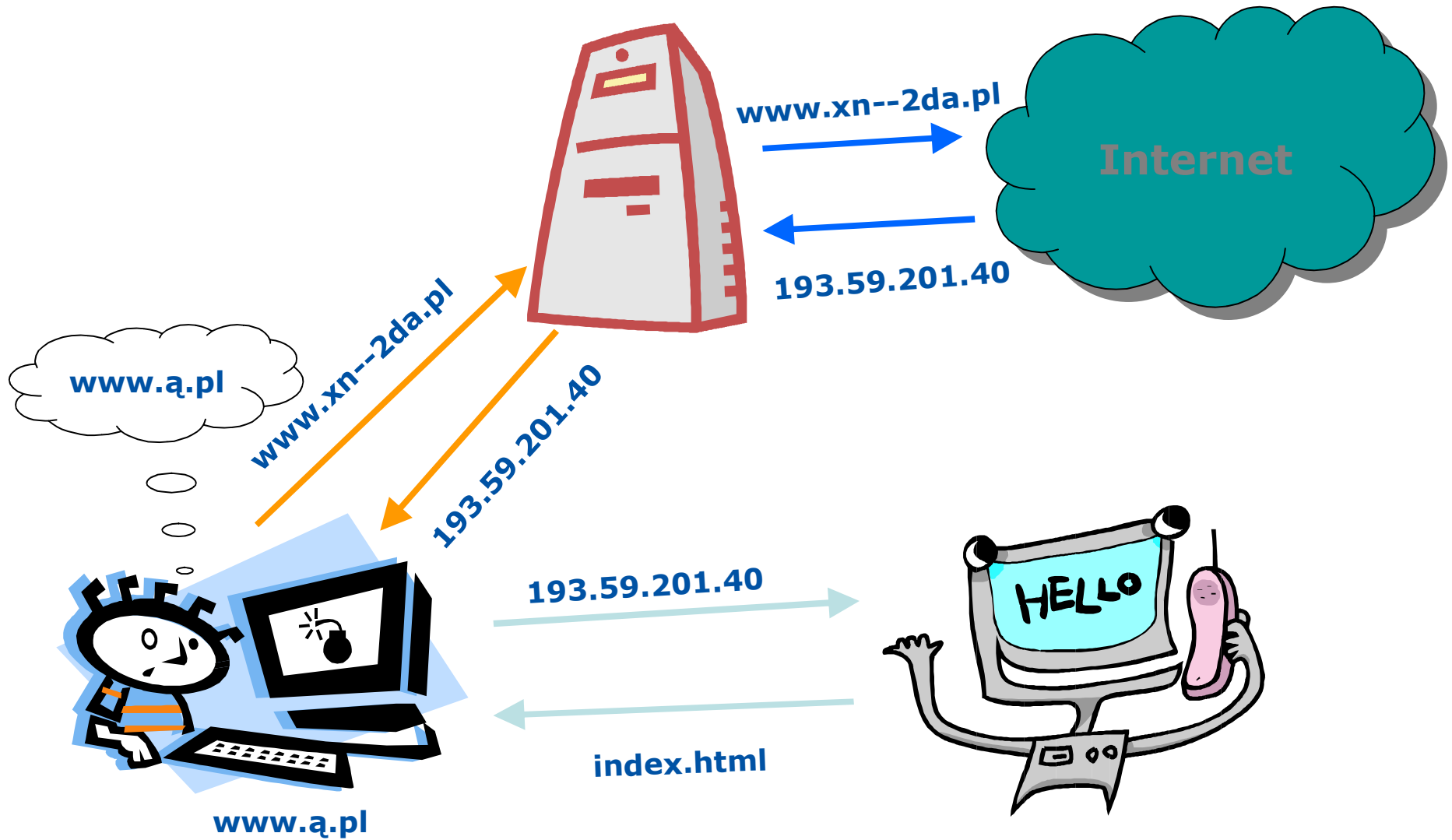
etykieta ACE

www.xn--ryczka-bxa01i.pl

Forma ACE domeny IDN przechowywana w zasobach DNSu.

ACE - ang. ASCII Compatible Encoding

Funkcjonowanie domen IDN



IETF

The Internet Engineering Task Force (www.ietf.org) - otwarta międzynarodowa społeczność internetowa, mająca na celu rozwój infrastruktury Internetu i dbająca o jego bezproblemowe funkcjonowanie. Społeczność tę stanowią projektanci sieci, operatorzy, badacze - naukowcy. Wynikiem pracy jednej z grup roboczych tej organizacji są dokumenty RFC 3490, 3491 i 3492 opisujące zasady funkcjonowania domen wielojęzycznych w Internecie. Dokumenty te zostały opublikowane jako RFC w marcu 2003 roku.

IANA

The Internet Assigned Numbers Authority (www.iana.org) - organizacja odpowiedzialna za wyznaczenie unikalnych wartości parametrów dla protokołów Internetu. To właśnie ta organizacja 14 lutego 2003 r. wyznaczyła prefiks „xn” dla domen IDN.

Dlaczego „xn--“?

Prefiks „xn” w 2003 roku został w drodze losowania wybrany jako międzynarodowe oznaczenie domeny która odzwierciedla nazwę wielojęzyczną.

Wcześniej funkcjonowały prefiksy „bq” i miały zastosowanie do testowych domen kodowanych innym algorytmem, który został przez IETF odrzucony (domeny z prefiksem “bq--” były obsługiwane jedynie przez przeglądarkę Opera)

Implementacja domen IDN w .pl

- 11 wrzesień 2003 – uruchomienie rejestracji domen idn z polskimi ogonkami ą (U+0105), ć (U+0107), ę (U+0119), ł (U+0142), ń (U+ 0144), ó (U+00F3), ś (U+015B), ź (U+017A), ż (U+017C)
 - 6 październik 2003 – rozszerzenie rejestracji domen IDN o niemieckie znaki diakrytyczne: ä (U+00E4), ö (U+00F6), ü (U+00FC)
 - 20 październik 2003 – dopuszczenie do rejestracji większości znaków z podzbiorów Unikodu Latin-1 Supplement oraz Latin Extended-A
 - 3 listopad 2003 – uruchomienie rejestracji domen IDN w skrypcie greckim, hebrajskim oraz arabskim.
 - 26 luty 2004 – uruchomienie rejestracji domen IDN w cyrylicy.
-

Dokumenty RFC

- “Internationalizing Domain Names In Applications (IDNA)” (RFC 3490) - jest to dokument definiujący domeny wielojęzyczne IDN oraz mechanizm zwany w skrócie IDNA, odpowiadający za implementację tego rodzaju domen.
- "Nameprep: A Stringprep Profile for Internationalized Domain Names (IDN)" (RFC 3491) - dokument ten jest profilem "Stringprepa", opisuje reguły przygotowania etykiet domen IDN.
- “Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications” (RFC 3492) - jest to dokument opisujący algorytm kodujący “Punycode” będący instancją algorytmu Bootstring. Algorytm ten pozwala jednoznacznie i odwracalnie przekodować ciągi znaków Unikodu w ciąg znaków ASCII składający się tylko z liter, cyfr i myślników.

oraz

- “Stringprep: Preparation of Internationalized Strings ("stringprep") (RFC 3454) – dokument opisujący ogólne zasady przygotowania tekstu unikodowego
-

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). RFC 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.

Główne funkcje Nameprepa

- Mapowanie (wykorzystane są tabele mapowania ze Stringprepa)

WIELKIE LITERY są mapowane na **małe litery** (tzw. case folding)

À ➔ à

- Normalizacja (NFKC)

Input 0061 0328

NFKC 0105

à | "a"+"ø"

- “Prohibition” (znaki zabronione np. spacje, znaki kontrolne)
- Test “bidi”

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.
-

IDNA

Internationalized Domain Names in Application

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 domeny 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.
-

Co „piszczący” w środku?

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.
-

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.zażółćgęślajażń.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.
-

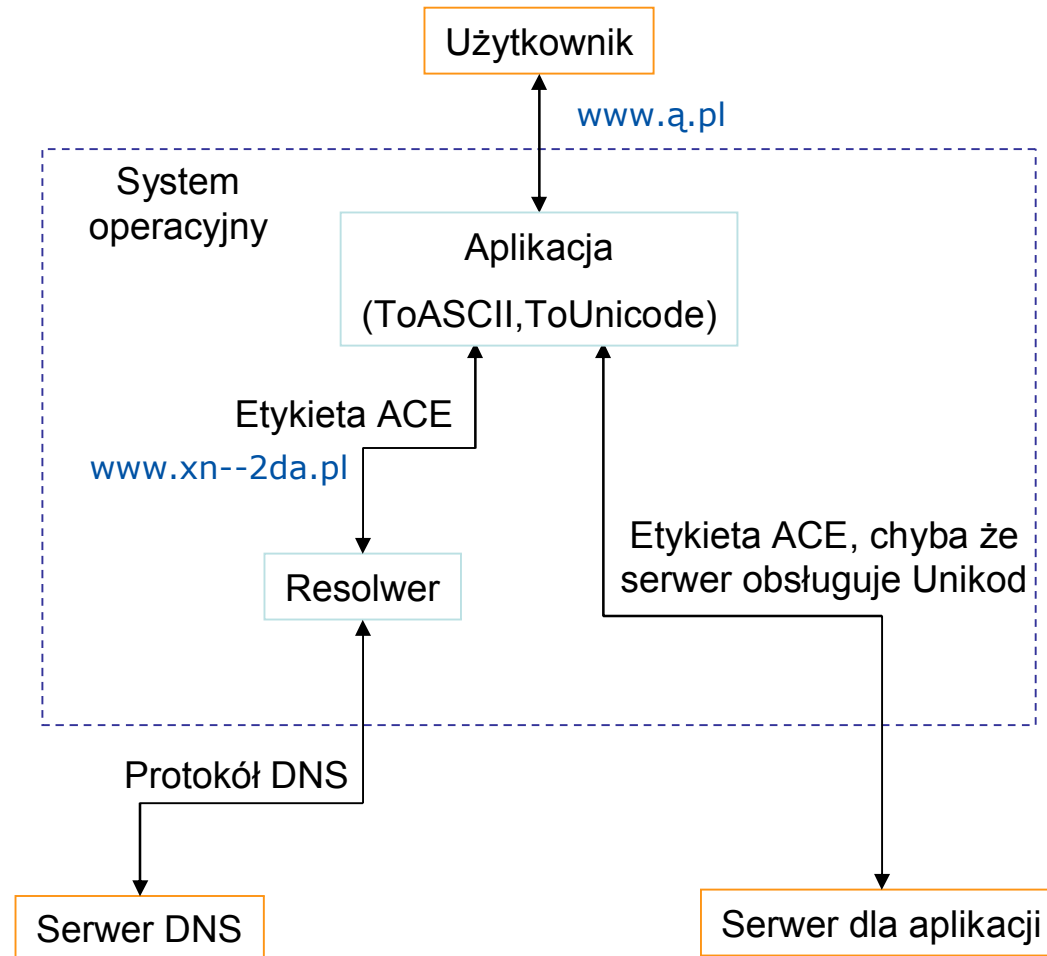
Operacja ToUnicode

- 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 na wejściu do tej operacji zostanie podana etykieta bez prefiksu ACE, to na wyjściu jest nie zmieniona etykieta.
 - Operacja ToUnicode nigdy nie zawodzi.
-

Jak działa operacja ToUnicode?

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

Schemat IDNA



Aspekty domen IDN

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.
-

Single-script spoofing

pepsi.pl

xn--e1avb8ff.pl

cocacola.pl

xn--l-7sbb7cbpbb.pl

paypal.pl

xn--payl-73d9g.pl

paypal.pl

xn--l-7sba6dbr.pl

paypal.pl

xn--apl-7cd1fta.pl

mbank.pl

xn--mbnk-63d.pl

Mixed-script spoofing

Latin script

Cyrillic script

Rejestry z idnami

- .jp (lipec 2003)
- .pl (11 wrzesień 2003)
- kr (październik 2003)
- .se (październik 2003)
- .dk (2 styczeń 2004)
- .museum (20 styczeń 2004)
- .no (9 luty 2004)
- .de (1 marzec 2004)
- .at (1 marzec 2004)
- .ch (1 marzec 2004)
- .lv (2004)
- .at (2004)
- .info (19 marzec 2004)
- .org (18 marzec 2005)
- .br (9 maj 2005)
- .gr (4 lipiec 2005)
- .fi (1 wrzesień, 2005)
- .cl (21 wrzesień 2005)
- .hu
- .ac
- .cn
- .io
- .li
- .lt
- .sh
- .th
- .tm
- .tw
- .vn

idnTLDs = 29

ccTLDs
+ gTLDs
<hr/>
= 264

www.polska.pl

www.πολωνία.pl

www.xn--kxae2aiif9d.pl

www.بولندا.pl

www.xn--mgbbv7fkk.pl

www.פולין.pl

www.xn--9dbiinv.pl

<http://www.dns.pl/IDN>

- Zasady rejestracji
 - Zbiory znaków dopuszczonych do rejestracji
 - Pogramy obsługujące domeny IDN
 - Narzędzia do konwersji domen IDN
 - Dokumenty RFC
-

Działania NASK w ITU

RESOLUTION 48

Internationalized domain names

(Florianópolis, 2004)

The World Telecommunication Standardization Assembly (Florianópolis, 2004),

recognizing

- a) relevant parts of Resolution 102 (Rev. Marrakesh, 2002) of the Plenipotentiary Conference;**
 - b) Resolution 133 (Marrakesh, 2002) of the Plenipotentiary Conference;**
 - c) relevant results of the first phase of the World Summit on the Information Society (WSIS);**
 - d) the evolving role of the World Telecommunication Standardization Assembly, as reflected in Resolution 122 (Marrakesh, 2002) of the Plenipotentiary Conference,**
-

considering

- a) that there needs to be an in-depth discussion of the political, economic and technical issues related to internationalized domain names (IDN) arising out of the interaction between national sovereignty and the need for international coordination and harmonization;
 - b) that intergovernmental organizations have had, and should continue to have, a facilitating role in the coordination of Internet-related public policy issues;
 - c) that international organizations have also had, and should continue to have, an important role in the development of Internet-related technical standards and relevant policies;
 - d) that the ITU Telecommunication Standardization Sector (ITU-T) has a record of successfully handling similar issues in a timely manner;
 - e) the ongoing activities of other relevant organizations,
-

instructs Study Group 17, in collaboration with other relevant study groups

to study IDN, and to continue to liaise and cooperate with appropriate entities in this area,

instructs the Director of the Telecommunication Standardization Bureau

to take appropriate action to facilitate the above and to report to the Council annually regarding the progress achieved in this area,

invites Member States

to contribute to these activities.

<http://www.itu.int/ITU-T/wtsa/resolutions.html>

Dotychczas prace koncentrowały się wokół 2 spotkań:

- ITU-T SG17 Moskwa, marzec 2005
- ITU-T SG17 Genewa, październik 2005

Na spotkanie w Moskwie zgłoszono i opracowano 13 kontrybucji, na spotkanie w Genewie 17 kontrybucji. Podczas spotkania w Moskwie odbyła się również specjalna sesja „tutorial”.

A.Bartosiewicz (NASK) pełni rolę „Rapporteur” dla tematyki IDN oraz przewodniczy spotkaniom Question 16.

Dokumenty ze spotkań Question 16 SG17 (kontrybucje oraz dokumenty wynikowe) można znaleźć pod adresem:

<http://www.itu.int/ITU-T/studygroups/com17/index.asp>

Following text of proposed Question on Internationalized Domain Names has been agreed during the IDN (proposed Q16/17) meeting October 7-11, 2005.

1 Motivation

The World Telecommunication Standardization Assembly (Florianópolis, 2004) in Resolution 48 instructed Study Group 17 (Security, languages and telecommunication software) to study Internationalized Domain Names (IDN). The belief is that IDN implementation will contribute to easier and greater use of the Internet in those countries where the native or official languages are not yet represented in ASCII characters.

2 Question

- a. What are the national, regional and international experiences of ITU Member States, ITU-T Sector Members and other relevant entities in the field of IDN?
 - b. What are the IDN needs of ITU Member States and Sector Members and how can those needs be addressed, taking into consideration the current IETF and ICANN work on IDN?
 - c. What telecommunication network standardization activity is required with regards to IDN, that may be needed in the form of ITU-T Recommendations or other ITU-T outputs?
-

3 Tasks

- a. Establish necessary liaison mechanisms with the relevant study groups and appropriate entities for this area of study.**
 - b. Develop a circular letter to identify the Member States and Sector Members issues and experiences with respect to IDN.**
 - c. Develop an analysis by which responses to circular letter could be categorized to identify issues and needs related to IDN, and gather experiences in the use of IDN**
 - d. In consultation with relevant entities develop a list of existing technical documentation stating the fundamentals of IDN to assist Member States and Sector Members in identifying their relevant issues and needs. This may include, but is not limited to:**
 - documentation relating to telecommunication network security risks accompanying implementation of IDN,**
 - issues regarding the use of regional language tables.**
 - e. Encourage contributions on multilingual issues relevant to IDN.**
-

- f. Identify deployment scenarios for IDN and options for assisting Member States and linguistic groups of Member States in their deployment actions.
- g. Provide documentation regarding languages currently standardized for deployment, those languages under standardization development and those not yet under consideration for standardization.
- h. Provide an annual progress report to assist the Director of TSB for use in his annual report to Council regarding activities on Resolution 48.

4 Relationships

- Questions: Q.5/17, Q.6/17, Q.7/17
 - Study Groups: SG 2 (Q.1/2)
 - Standardization organisations: IETF, ISO/IEC
 - Other organisations: ICANN, UNICODE Consortium (pending recognition by TSB in accordance with ITU-T Recommendation A.5)
-

Jak uczestniczyć w pracach ITU w zakresie IDN?

- Uczestnictwo w spotkaniach SG17 Q16 (dwukrotnie w ciągu roku)
 - Analiza nadesłanych i wysyłanie własnych kontrybucji
 - Lista dyskusyjna Question 16
-