

Title: Experiences on Internationalised Domain Names implementation under .pl

Authors: A. Bartosiewicz, K.Olesik

Purpose: EURID IDN Advisory Board conference call

Abstract: This document presents the way the Internationalized Domain Names were introduced under .pl.

NASK began his work on Internationalised Domain Names (IDNs) in 2002. A work plan included:

- a) a study of the subject;
- b) test registrations;
- c) establishment of a registration policy;
- d) preparation of an on-line conversion tool and a registration system;
- e) trainings of a contact centre;
- f) creation of an IDN WWW sub-service (<http://www.dns.pl/IDN/>) and
- g) PR activities.

Preparations for the launch of the IDN registration were based on the following assumptions:

- a) no Sunrise period which entailed First Comes First Served (FCFS) rule;
- b) no trademark protection;
- c) only the ACE form of an IDN were accepted for registration; a registrant was supposed to prepare his Ace form of desired IDN by himself;
- d) no language tags;
- e) no automatic variant/equivalent domain name registration;
- f) registration of domain names with other prefixes prohibited.

Aforementioned assumptions entailed minimum changes to the NASK Registry System which required only implementation of the IDN validation mechanism - ToUnicode and ToASCII operations plus character filter. Such approach had no impact on the billing system, whois service and online electronic forms. Because the subject of the agreement between registry and registrant was ACE form of an IDN, the Terms and Conditions had not to be changed. To summarize, a domain name with "xn-" prefix was present on invoices, whois, registration applications, domain delegation change requests, etc.

The launch of IDN was preceded by press conference (March 2003) and series of presentations on international conferences: CENTR Administrative Workshop (Frankfurt 2003), CENTR Technical Workshop (Frankfurt 2003), IETF (Vienna 2003) and RIPE (Amsterdam 2003). On 11 August 2003 the rules of registering IDNs under .pl were published at IETF as internet draft. All changes in IDN registration rules had been reflected in the subsequent versions of this internet draft. Nowadays, up-to-date registration rules are available at www.dns.pl/IDN only.

The registration of internationalised domain names (IDNs) under .pl launched on 11 September 2003 and officially announced on National Telecommunication Symposium at the same day. Initially, an IDN could contain only Polish diacritics, namely: 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). The launch of Polish diacritics was divided into two stages to reduce likelihood of high server load and ensure stable work of the registry system. First stage, 11 September 2003 - registration of IDNs under .pl

Contact: Krzysztof Olesik, Andrzej Bartosiewicz
NASK
Poland

Email: kolesik@NASK.pl
andrzejb@NASK.pl

zone only; second stage, 18 September 2003 - registration under second level domain names governed by NASK.

Number of registration during the first stage was as follows (registration started at 6 a.m.):

- 350 at 6:30 a.m.;
- 704 at 11:11 a.m.;
- 766 at 12:00 a.m.;

where average daily registration, before the IDNs launch, was at a level of 250 normal domains a day. Figures depicting some statistics are enclosed in appendix A.

The character collection allowed in IDNs was extended according to the following schedule:

- a) 6 October 2003 - adding the German diacritics ä (U+00E4), ö (U+00F6), ü (U+00FC).
- b) 20 October 2003 - adding selected characters from Unicode scripts: Latin-1 Supplement and Latin Extended-A;
- c) 3 November 2003 - release selected characters from Unicode scripts: Arabic¹, Greek, Hebrew;
- d) 26 February 2004 - release of characters derived from Unicode Cyrillic script.

Each character to be added to actual collection was analyzed as to whether:

- a) it is a small letter or a digit (ligatures were exception to the rule, e.g. U+0153);
- b) it is at least used in one of European languages (only Arabic was an exception);
- c) an input and an output of a normalization process is the same character; if the output of normalization was different than the input then the new character from the output (not present yet in the collection) was added to the character collection. (It will be explained below.)

NASK accepts for registration only proper ACE form of an IDN. Therefore, it means that each character encoded (by means of Punycode algorithm) in the ACE form had been normalized first. For example, if one want to register an IDN containing character “ł” (U+0140 LATIN SMALL LETTER L WITH MIDDLE DOT) then as a result of the normalization process in the ACE form will be encoded the following string “l” (U+006C LATIN SMALL LETTER L) followed by “.” (U+00B7 MIDDLE DOT) instead of “ł”. This impose that “middle dot” have to be included in allowed character set. Now, considering “ij” character (U+0133 LATIN SMALL LIGATURE IJ), we do not need to add this one to allowed character set because it is normalized to l (U+006C LATIN SMALL LETTER L) followed by j (U+006A LATIN SMALL LETTER J). To conclude, NASK precisely determine characters which can be encoded in the ACE form of an IDN.

The full list of characters is available at http://www.dns.pl/IDN/allowed_character_sets.pdf. The NASK’s language tables published at IANA language tables registry (March 2004) may be of interest as well.

Since the release of the Cyrillic script, allowed characters have been divided into several sets² in order to prevent mixing of characters derived from different scripts. It means that valid IDN (ACE form), after ToUnicode operation was applied to, must not contain characters being derived from different sets. Such solution eliminates possibilities of “mix-script spoofing” (see UTR 36). Nevertheless, this approach do not prevent threat of “single-script spoofing”. To address this problem, “ASCII-like” characters have been selected from the Cyrillic set, i.e.: a (U+0430), e

¹ Arabic set is currently suspended.

² see up-to-date IDN registration policy and definition of character sets available at http://www.dns.pl/IDN/idn-registration_policy.txt

(U+0435), o (U+043E), p (U+0440), c (U+0441), y (U+0443), x (U+0445), s (U+0455), i (U+0456), j (U+0458). Any IDN consisting only from hyphen “-”, digits 0-9 and mentioned above characters has to undergo a check-up before a registration. Let’s consider Cyrillic IDN 123-peace.pl (xn--123--83d3ab9fl.pl). All Cyrillic characters in the label belong to the set of ASCII-like characters. The ASCII-like characters of the label are mapped to their similar ASCII characters and next received in this way ASCII variant of the Cyrillic IDN is checked against registered or booked domain names in the registry system. If there is no such domain, then the registration process may be continued. This solution do not protect the Cyrillic IDNs already registered, i.e. if one wants to register 123-peace.pl (not IDN), then no check-up is performed as to whether a Cyrillic variant of the domain exists already. This approach was devised to protect merely the ASCII domains from spoofing.

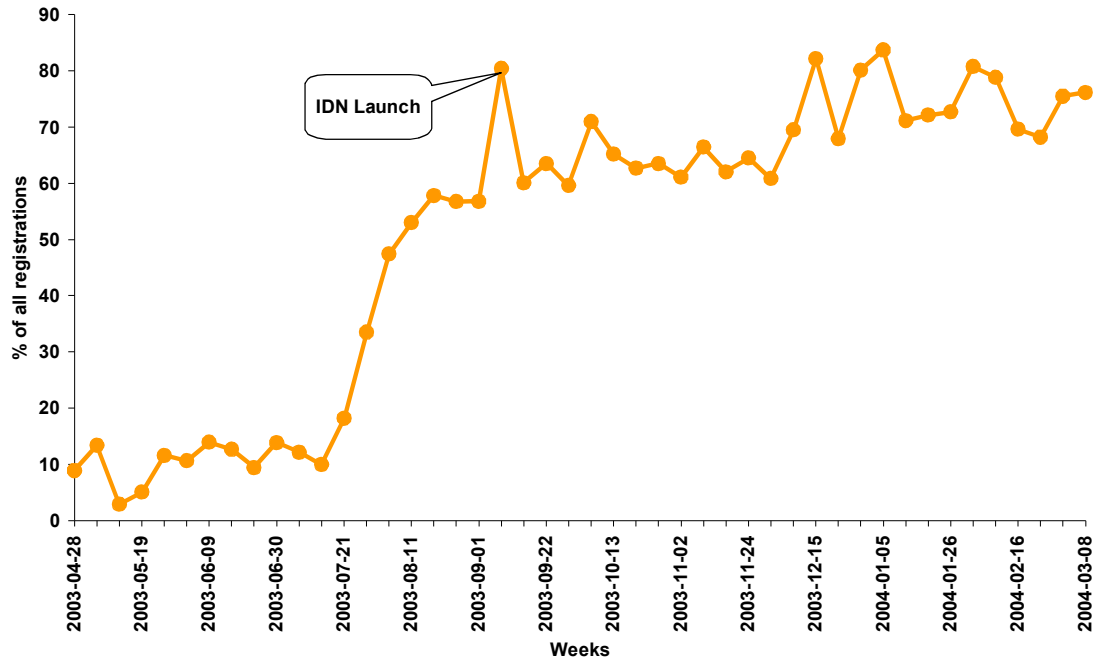
At present is being used more complex and comprehensive preventive system against spoofing. Not only ASCII-like characters from Cyrillic script are considered but characters being derived form other scripts enabled under .pl as well. This special “anti-homograph” mechanism is implemented in the registry system. The mechanism is to filter out all single-script spoofable domains (not only IDNs) during the registration process, create variant domains in accordance with a mapping table and check against existing domains (both IDNs and ASCII one). A registration process will be finished successfully when neither being registered domain nor its possible variants exist in the registry system. The algorithm and mappings are explained in the IDN Registration Policy available at <http://www.dns.pl/IDN/idn-registration-policy.txt>.

Conclusion:

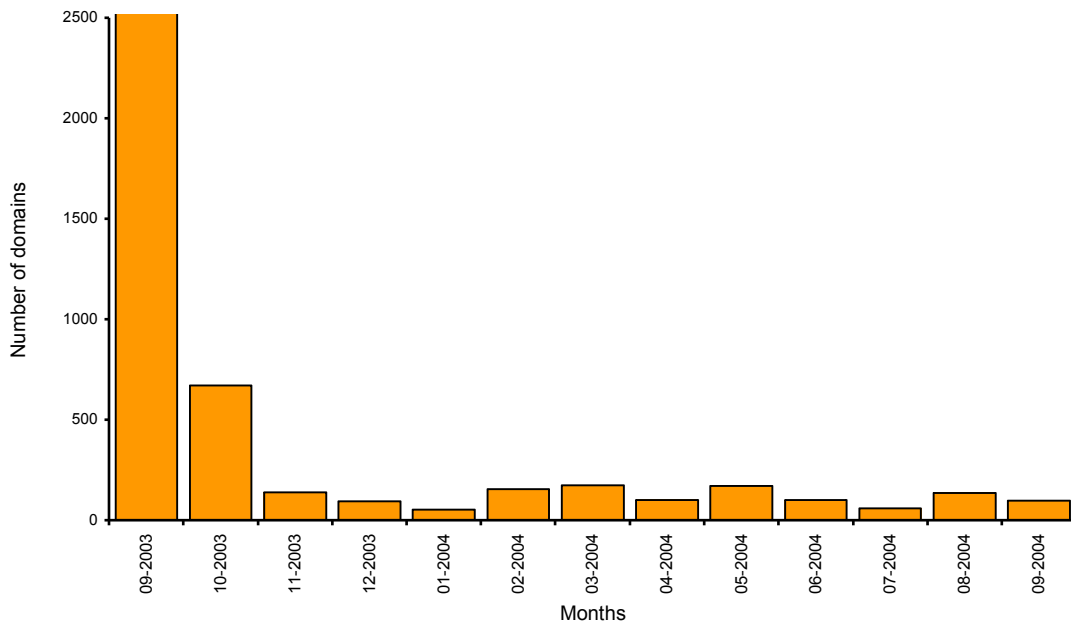
A technical introduction of IDNs is not so problematic as establishment of a good IDN registration policy. Technical essentials of IDNs are well-known and described in many papers whereas relevant policy issues, like homographs and spoofing issue, are still under discussion. A registry may choose a language-based approach to IDN registration or script-based one. Which is better? The answer to this question is not so easy because it depends on a “profile” of a registry - whether the registry wants to enable one or just a few languages or a whole bunch of languages, e.g. for particular geographic region; and whether the languages belong to one or different writing systems or else Unicode scripts. One is sure, that creation of a language tables is very complex matter and requires good knowledge of a particular language. A registry may try and build their own language/script tables or take an advantage of experiences other registries.

Appendix A

1) Percentage increase in weekly registrations of all domain names.



2) Monthly registration of IDNs



3) weekly registration of IDNs

