

ITU and UNESCO Global Symposium on Promoting the Multilingual Internet

Geneva, 9-11 May 2006

Paper Submitted by: NASK, Poland
Krzysztof Olesik
Andrzej Bartosiewicz

Title: Experiences on implementation of Internationalised Domain Names under .pl

For Discussion in Session: 6 - Thematic Session on Advancing the Deployment of IDN

Abstract

This Contribution is intended to present a way the Internationalized Domain Names (IDNs) has been introduced under .pl - polish ccTLD (country code Top Level Domain) governed by NASK (Research and Academic Computer Network). To learn more about NASK and its activity go to http://www.nask.pl/run/n/Who_we_are. This document gives an overview of actions taken in a preparation period before an IDN launch at NASK, first assumptions and rules of IDN registrations, development of an IDN registration policy and statistics.

Preparations

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 an ACE form of an IDN were accepted for registration; a registrant was supposed to prepare his Ace form of the 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 an IDN validation mechanism - ToUnicode and ToASCII operations plus character filter. Such approach had no impact on a billing system, a whois service and online electronic forms at NASK. Because a subject of an agreement between registry and registrant is an 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

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

The registration of internationalised domain names (IDNs) under .pl launched on 11 September 2003 and was officially announced on National Telecommunication Symposium on 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 a high server load and ensure stable work of the registry system. The first stage, 11 September 2003 - registration of IDNs under .pl zone only; the 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 from years 2003/2004 are enclosed in appendix A.

Nowadays there is more than 4600 IDNs registered under .pl. In 2005, average registration of IDNs was at a level of 150 domains per month which account for 1% of the total monthly registrations. Figures presenting statistics from year 2005 are attached in appendix B. Most of registered IDNs comprises of characters being derived form a latin set¹.

Policy development

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: Arabic², Greek, Hebrew;

¹ allowed characters for IDN registration are divided into several sets, see IDN registration policy at <http://www.dns.pl/IDN/idn-registration-policy.txt>

² Arabic set is currently suspended.

- 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 are the same character; if an output of a normalization process was different than an input then a new character from an output (not present yet in the collection) was added to the character collection. (It will be explained below.)

NASK accepts for registration only a proper ACE form of an IDN. Therefore, it means that each character encoded (by means of Punycode algorithm) in an ACE form had been normalized first. For example, if one want to register an IDN containing character “ḷ” (U+0140 LATIN SMALL LETTER L WITH MIDDLE DOT) then, as a result of a normalization process, in the ACE form will be encoded the following string “ḷ” (U+06C LATIN SMALL LETTER L) followed by “.” (U+00B7 MIDDLE DOT) instead of “ḷ”. This impose that “middle dot” have to be included in the 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 “ḷ” (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 an ACE form of an IDN.

The full list of allowed characters is available at <http://www.dns.pl/IDN>. It has also been published as Contribution C 31 “Experience of IDN implementation in Poland” at ITU-T Study Group 17 (for information on registration policy see also Contribution C 12 “Internationalized Domain Names in Poland”). The NASK’s language tables published at IANA language tables registry (March 2004) may be of interest as well (see <http://www.iana.org/assignments/idn/registered.htm>).

Since the release of the Cyrillic script, allowed characters have been divided into several sets³ in order to prevent mixing of characters derived from different scripts. It means that a 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, <http://www.unicode.org/reports/tr36/>). 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 (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 ASCII domains from spoofing.

³ see up-to-date IDN registration policy and definition of character sets available at <http://www.dns.pl/IDN/idn-registration-policy.txt>

Plans

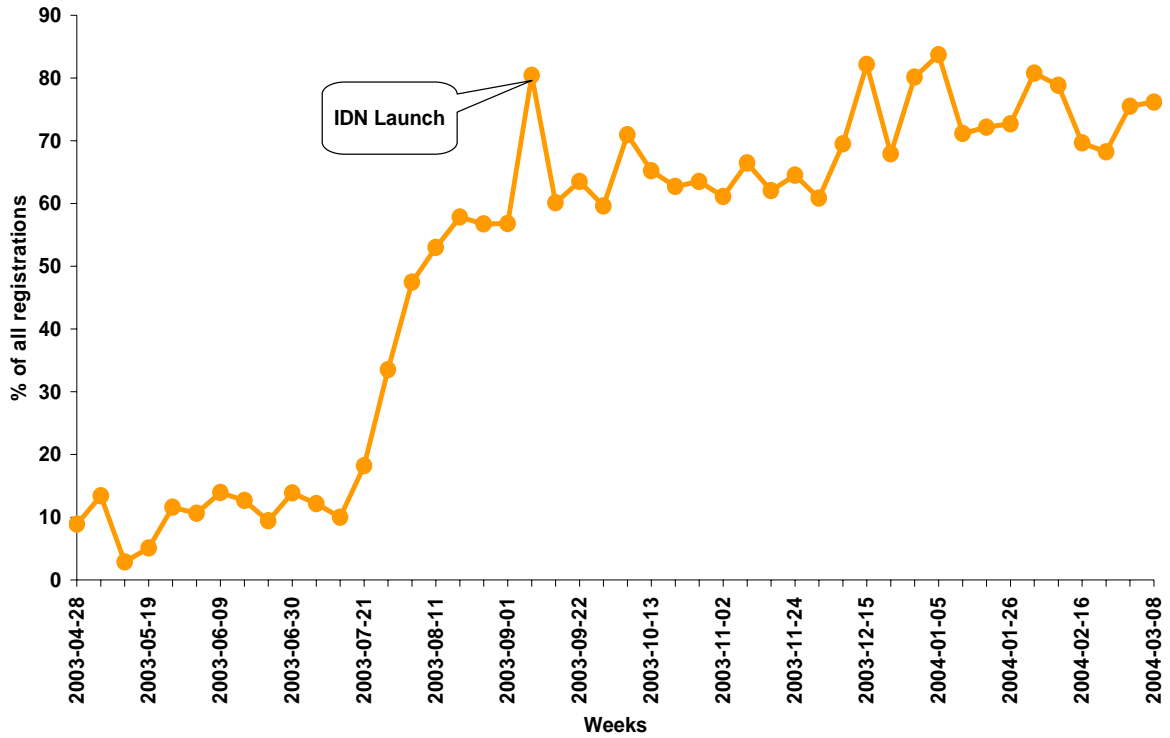
At present NASK is about to introduce more complex and comprehensive preventive system against spoofing. Not only ASCII-like characters from Cyrillic script are considered but characters being derived from other scripts enabled under .pl as well. The special “anti-spoofing” mechanism will be implemented in the registry system. The mechanism is to filter out all single-script spoofable domains (not only IDNs) during a 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.

Conclusion

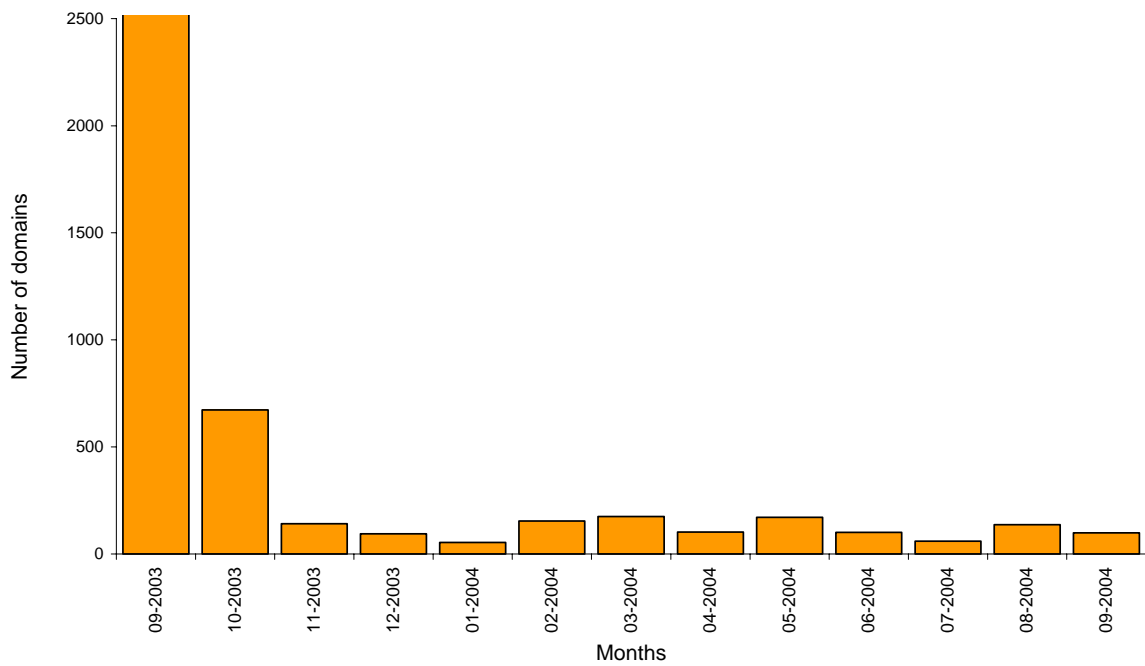
A technical introduction of IDNs is not so problematic as establishment of a good IDN registration policy. Technical essentials of IDNs are well-known and described in many papers whereas relevant policy issues, like “variant character” and spoofing issue, are still under discussion. A registry may choose 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 a 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 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 advantage of experience of other registries.

Appendix A

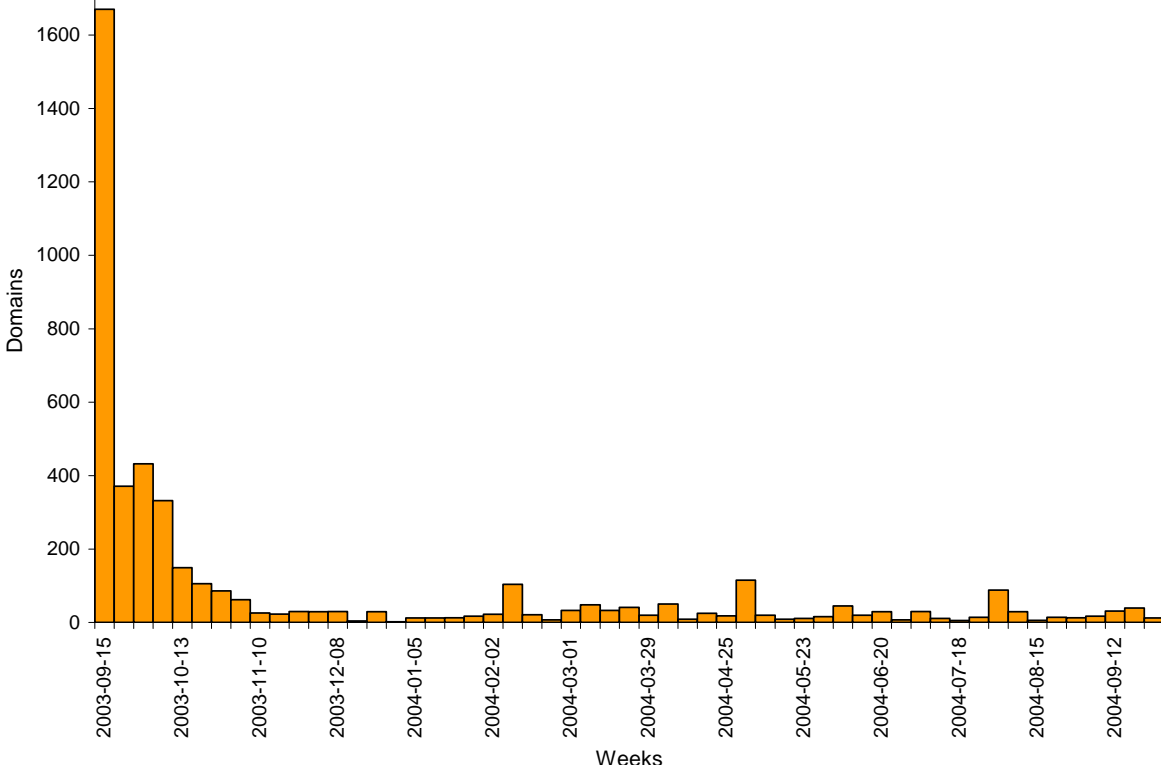
1) Percentage increase in weekly registrations of all domain names (period: 04.2003-03.2004)



2) Monthly registration of IDNs (period: 09.2003 - 09-2004)

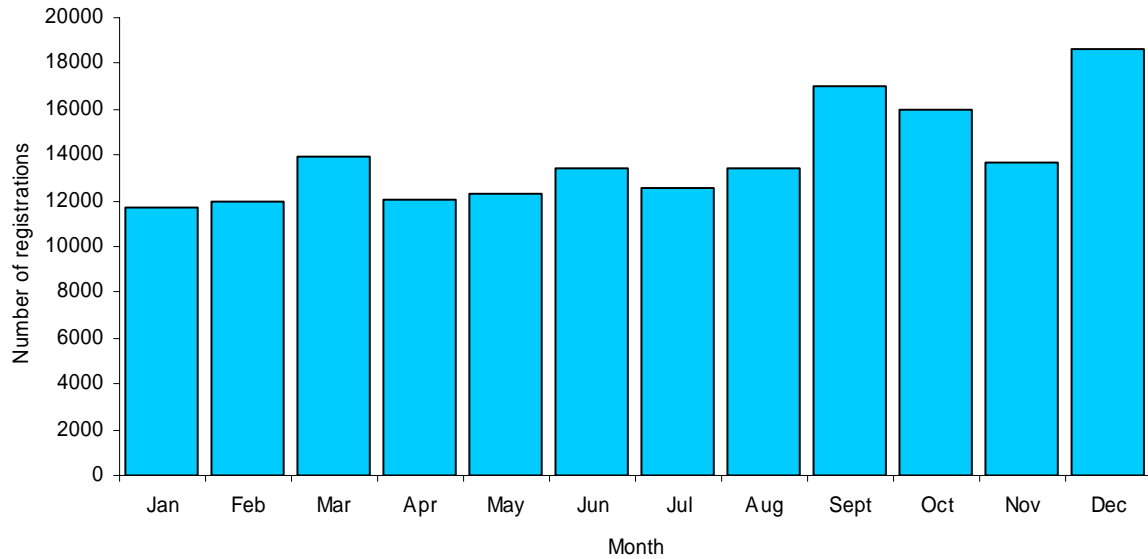


3) Weekly registration of IDNs (period: 09.2003 - 09-2004)

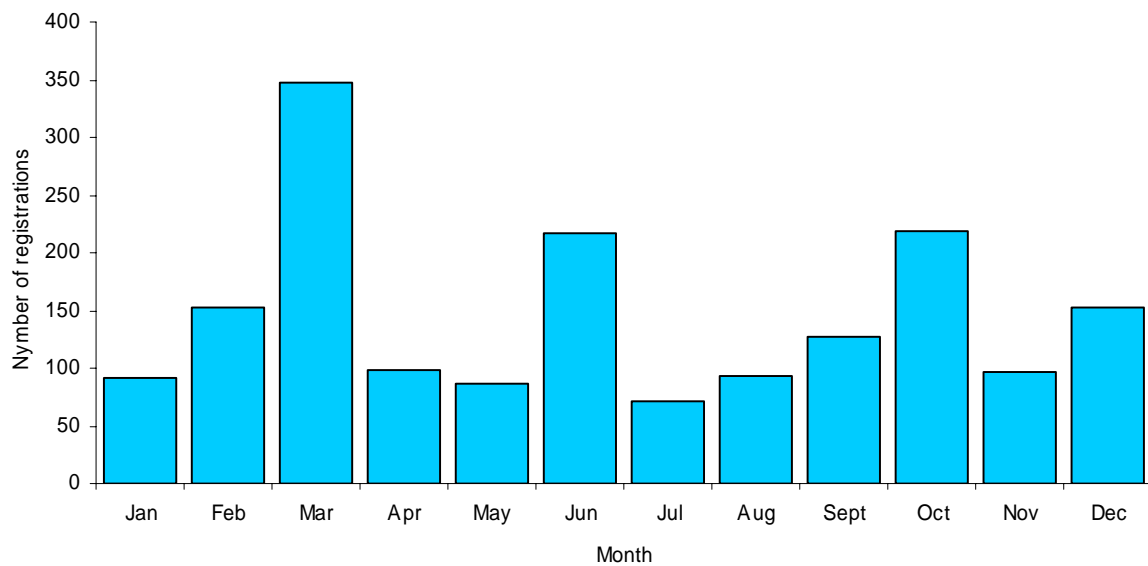


Appendix B

1) Monthly registration of all domain names in 2005



2) Monthly registration of IDNs in 2005



3) Number of IDNs per month

