Breaking the Linguistic Barriers to Access the Internet

Sarmad Hussain, Jean-Jacques Sahel, Gabriella Schittek

Internet Corporation for Assigned Names and Nunbers (ICANN)

6-Rond-Point Schuman
B-1040 Brussels, Belgium

{sarmad.hussain, jean-jacques.sahel, gabriella.schittek}@icann.org

Abstract

This paper explains how Internet's Domain Name System (DNS) is expanding so that it now allows for using domain name suffixes (called top-level domains) to be represented in different scripts used around the world. This expansion allowing domain names in local languages (called Internationalised Domain Names) makes it easier for the majority of the world's population to have an equal opportunity to access the Internet. The paper also examines the technical hurdles this expansion is facing - the so-called Universal Acceptance issue for domain names, what ICANN and the community is doing to address this and how others, including the academic community worldwide, can contribute to improving the multilingual access to the internet.

Keywords: Domain Name System, Top-Level domains, ICANN, Internationalised Domain Names, Universal Acceptance

1. Introduction

Like the physical world, one needs an address to get around in the online world. However, as the online world or the Internet is composed of a network of computers and other devices connected together, their addresses are represented as sequences of numbers which these machines can understand. These are called Internet Protocol or IP addresses and are assigned to the devices which are online. Though numbers are native to machines, these are not easy for the humans to remember. Therefore, to help humans navigate the web of online information, the Internet's Domain Name System (DNS) is used. The DNS enables to assign domain names for the IP addresses. For example, 192.0.2.0 is an IP address and "example.com" is a domain name. Clearly, the domain name is easier for the humans to remember.

The Internet Corporation for Assigned Names and Numbers (ICANN) is the organization tasked with ensuring that the Internet's names and numbers known as 'unique identifiers', which are used by the DNS, work smoothly. ICANN is a global organisation, which operates according to a so called Multistakeholder Model, meaning that everyone interested in the DNS – whether from government, business, the technical community, academia or civil society - can contribute to ICANN's work on policy development around these functions. These include policies developed around Top Level Domains (TLDs), such as .com, or .org, which ICANN is responsible for adding to the root zone directory, to make them accessible on the Internet.

2. Internationalised Domain Names

Initially, the DNS was developed using the LDH scheme, allowing domain names only using Letters (A-Z, a-z), Digits (0-9) and Hyphen (-). However, as the Internet evolved into a global means of exchanging information and communicating, with communities across the globe connecting through it, the LDH scheme was found to be too limited and created linguistic barriers, because the populations around the world used different scripts and writing systems natively and, in some cases, were not familiar with the characters used in LDH scheme, which were limited to the Latin script. The global linguistic diversity required that the DNS be internationalized, by allowing it to provide multilingual support. Work on Internationalized Domain Names (IDNs) started in the late 1990s and the Internet community developed a standard to allow IDNs based on the Unicode standard. This standard, called IDN in Applications, was published by standards body the Internet Engineering Task Force (IETF) first as IDNA2003 and then revised to IDNA2008. This allowed people around the world to represent domain names in local languages and scripts.

The DNS has been based on ASCII code, which originally consisted of 128 characters. Even from these, only 63 characters were allowed through the LDH scheme for domain names. IDNA2008 expanded the base character set to span the Unicode standard which has a much larger repertoire. Unicode version 11.0 released in June 2018 contains support for 146 scripts and 137,439 characters. Based on the algorithm defined by the IDNA2008 standard, a smaller subset of these characters has been allowed, mostly to include letters, marks and digits from these scripts (excluding characters for punctuation and symbols). This significantly expands the number of characters which could be permissible for domain names.

3. Creating Internationalised Domain Names

3.1 Defining the Character Repertoire

ICANN, entrusted by the global Internet community with the responsibility to ensure the stable and secure operation of the Internet's unique identifier systems, needs to define the relevant character repertoire and

additional constraints for the names which can be allowed in the Internet's root zone, the TLDs, as stipulated by IDNA2008. This work also needs to take into consideration the RFC 1123 standard by IETF, which further requires the TLDs to be alphabetic.

ICANN has undertaken a programme to define these constraints to determine valid and unique TLDs for all the scripts used globally. This requires analysis of each script to determine: (i) the alphabetic character repertoire for representing domain names; (ii) the characters which are considered "same" or interchangeable by the relevant script community; and (iii) any additional constraints to determine which label can be a valid TLD. The ICANN community understands that this expertise lies with the communities which use the script. Therefore, this work is being carried out by different community-based panels for each script, which consist of linguistics, technology and policy experts. The proposals developed by these community-based panels are reviewed and integrated by an independent panel of experts in linguistics, Unicode and the DNS.

Following the principles set out in the RFC 6912 standard and recommendations in the procedure[1] finalized by the community, each panel first determines the minimal list of code points in the Unicode needed for representing TLDs in the relevant script. As all major languages which actively use the script have to be covered, a reasonably comprehensive analysis needs to be undertaken for this purpose. For example, as the authors of the proposal[2] for Devanagari script note that though Nukta (a dot below) is normally written with consonants, it is included with some specific vowels as well, due to use in Santali language, such as आ (Unicode code points: U+0906 and U+093C). They also disallow the character ऌ (U+090C) from use for Devanagari TLDs because it is not in modern usage. Similar analyses are being conducted by different panels for 28 scripts identified for use in TLDs in the first phase of the work[3] based on modern and everyday use of these scripts. In addition to determining what is the right alphabetic subset of characters allowed for forming TLDs in a script, the panels are also determining characters which should be variants within or across different scripts. This is important to keep the DNS identifiers "unique" from the perspective of the users, helping prevent some undesirable consequences: the (mis)use of domain names in fraud such as 'phishing' cybercrime attacks, whereby the user is intentionally confused into believing that they are visiting a genuine website, for example, when in fact they are directed to a fake website with domain name perceived as "same" by the end user. The table below presents some examples of variant characters, between Latin and Cyrillic scripts and between Telugu and Kannada scripts, as being suggested by the relevant community-based panels[4].

Latin and Cyrillic Script Variant Characters		Telugu and Kannada Script Variant Characters	
a (U+0061)	а (U+0430)	అ (U+0C05)	ಅ (U+0C85)
c (U+0063)	с (U+0441)	ఆ (U+0C06)	ಆ (U+0C86)
e (U+0065)	е (U+0435)	ఇ (U+0C07)	ಇ (U+0C87)
o (U+006F)	о (U+043E)	ఒ (U+0C12)	ಒ (U+0C92)
i (U+0069)	і (U+0456)	ఘ (U+0C1C)	ಘ (U+0C9C)

Table 1. Examples of variant code points

3.2 Introducing Internationalised Top Level Domains into the Root Zone

This work allows for determining valid IDN TLDs and their IDN variant TLD labels based on the data and the associated algorithm produced by the community, thus providing a transparent multi-stakeholder mechanism to allow for secure and stable multilingual domain names for the root zone. In consultation with the community, ICANN aims to integrate the LGR for the root zone into the future applications for generic and country code TLDs (IDN gTLDs and ccTLDs). With the IDN gTLDs and ccTLDs available, it allows users around the globe to have access to, create and exchange on the Internet in their own languages and scripts, allowing them to claim (and use) a domain name that best reflects their identity. IDNs contribute towards bringing the next billion people online and growing the global Internet economy.

However, to ensure a fully satisfying user experience of IDNs on the Internet, there is still plenty of work to be done.

4. Universal Acceptance (UA) of Domain Names and Email Addresses

Since 2010, when the DNS expanded with the inclusion of IDN TLDs, first by the introduction of country codes such as .中国 (China), .срб (Serbia) or . ලංකා (Lanka) and a few years later with generic TLDs, such as .グーグル (google), .我爱你 (Iloveyou), .닷넷 (net), a new problem arose.

Many organizations and businesses are not prepared for this expansion and have not updated their online systems to universally accept new domains. They may have data validation routines embedded in their code that are out of date, or may still be operating in an ASCII only world, limited to Latin script.

As a result, many applications, Internet-connected devices and systems are unable to accept, validate, store, process or display all domain names or internationalized email addresses. For instance, because they do not recognise them as valid, email applications may not

[1] https://www.icann.org/en/system/files/files/lgr-procedure-20mar13-en.pdf

[2] https://www.icann.org/en/system/files/files/proposal-devanagari-lgr-27jul18-en.pdf

[3] https://www.icann.org/en/system/files/files/msr-4-overview-25jan19-en.pdf

[4] Completed proposals by the panels available at https://www.icann.org/resources/pages/lgr-proposals-2015-12-01-en

accept internationalized email addresses, online forms reject the internationalized email address, or online portals recognize the domain names which use new TLDs, which may be longer than four letters, like .pizza or .movie.

All this not only results in a poor user experience, but more importantly, leads to a lower usage of newer or IDN domain names than intended. It is just simpler for the end-user to avoid problems by continuing using the well-known TLDs like .com and .org. As a consequence, millions of people around the world can't fully experience the benefits of the Internet in the language they speak, or the domain name or email they have regsitered – just because applications are not yet UA ready.

4.2 Encouraging UA Readiness

The incorporation of these new domains across the global Internet is not an entirely automatic process. However, the efforts required for software and application owners for UA are not particularly complex either.

ICANN has therefore, together with community including the industry leaders including Apple, GoDaddy, Google, Microsoft and Verisign, has developed a Universal Acceptance Steering Group (UASG). The UASG exists to help stakeholders ensure their systems are UA-ready and able to accept all domain names and email addresses. The Steering Group is also working on communicating the issue and solutions to CIOs, web administrators, application developers and others who have an important role to play in making sure their applications. Through dedicated meetings, publications and speaking opportunities at relevant conferences, the UASG is trying to reach the relevant audience. A lot of efforts are also directed at academia, to make the next generation of programmers and web developers aware of the importance of Universal Acceptance of domain names and email addresses and how to implement it. For this purpose, an entire curriculum has been developed, which is offered to professors and teachers to be used in their lectures. Plenty of other material has also been developed to support the stakeholders in updating their systems, which is available at www.UASG.tech.

5. Conclusion

With the Internet having crossed four billion Internet users[5] – most coming from parts of the world not using Latin script - it is important that every end-user gets an equal opportunity to use and explore the Internet in their script, and not held back by technical restrictions, which can easily be fixed.

Therefore, it is important that industry leaders, academia and the technical community understand the value and significance of becoming Universal Acceptance ready and choose to address the issue appropriately. The best way in making this happen, is ensuring that the entire Internet community and end-users are aware and ready to inform and put the appropriate pressure on the industry. Academia can play a major role in this endeavour, also in training the next generation of technologists to be ready and supportive of the global, ever-growing and the multilingual Internet.

[5] https://www.internetworldstats.com/stats.htm

References

Sullivan, A., Klensin, J. and Kolkman O. (2013). *Principles for Unicode Code Point Inclusion in Labels in the DNS*. Internet Engineering Task Force. Retrieved from: https://tools.ietf.org/html/rfc6912. Access date: April 24, 2019.

Davies, K. and Freytag, A. (2016). *Representing Label Generation Rulesets Using XML*. Internet Engineering Task Force. Retrieved from: https://tools.ietf.org/html/rfc7940. Access date: April 24, 2019.