Glossary IDNs

Internationalised Domain Names

In an attempt to ensure that discussions regarding IDNs take place in a consistent manner ICANN has published an IDN Glossary. The glossary terms can be used freely and is expected to be expand over time. If you have suggestions for additions and/or changes to the glossary please submit these to idn-glossary@icann.org. Comments will be posted publicly in the discussion forum at http://forum.icann.org/lists/idn-glossary/.

Historically, domain names on the Internet were restricted to using a limited set of ASCII characters (i.e. a-z, 0-9 and "-"). However, with the increasing use of the Internet in all regions and by diverse linguistic groups of the world, the demand for multilingual domain names has become more intense. Various acronyms are used widely in communications around internationalising the domain name space. Explanations for many of these acronyms are provided below to help make this topic simpler to understand.

A

ACE (ASCII Compatible Encoding)

ACE is a system for encoding Unicode so each character can be transmitted using only a limited set of ASCII characters (i.e. a-z, 0-9 and "-"). This is used because applications that use the DNS protocol may not reliably handle other values.

ASCII (American Standard Code for Information Interchange)

ASCII is a common numerical code for computers and other devices that work with text. Computers can only understand numbers, so an ASCII code is the numerical representation of a character such as 'a' or '@'. When mentioned in relation to domain names or strings, ASCII refers to the fact that before internationalisation only the letters a-z, digits 0-9, and the hyphen "-", were allowed in domain names.

C

Character

For the purposes of discussing IDNs, a "character" can best be seen as the basic graphic unit of a writing system, which is a script plus a set of rules determining how it is used for representing a specific language. However, domain labels do not convey any intrinsic information about the language with which they are intended to be associated, although they do reveal the script on which they are based. This language dependency can unfortunately not be eliminated by restricting the definition to script because in several cases (see examples below) languages that share the same script differ in the way they regard its individual elements. The term character can therefore not be defined independently of the context in which it is used.

In phonetically based writing systems, a character is typically a letter or represents a syllable, and in ideographic systems (or alternatively, pictographic or logographic systems) a character may represent a concept or word.

The following examples are intended to illustrate that the definition of a character is at least two-fold, one being a linguistic base unit and the other is the associated code point.

U-label 酒: Jiu; the Chinese word for ‘alcoholic beverage’; Unicode code point is U+9152 (also referred to as: CJK UNIFIED IDEOGRAPH-9152); A-label is xn—jj4

U-label 北京: the Chinese word for ‘Beijing’, Unicode codepoints are U+5300 U+4EAC; A-label is xn—1lq90i

U-label 東京: Japanese word for ‘Tokyo’, the Unicode code points are U+6771 U+4EAC; A-label is xn—1lqs71d

U-label ایکوم; Farsi acronym for ICOM, Unicode code points are U+0627 U+06CC U+0643 U+0648 U+0645; A-label is xn—mgbodgl27d.

D

DNS (Domain Name System)

The DNS makes using the Internet easier by allowing a familiar string of letters (the “domain name”) to be used instead of the arcane IP address. So instead of typing 207.151.159.3, you can type www.internic.net.

I

IDNA (Internationalised Domain Names in Application)

IDNA is a protocol defined in RFC 3490 by the Internet Engineering Task Force (http://www.ietf.org) that makes it possible for applications to handle domain names with non-ASCII characters. IDNA converts domain name strings with non-ASCII characters to ASCII domain name labels that applications that use the DNS can accurately understand. Not all characters used in the world’s languages will be available for use in domain names. Hence IDNA is not able to convert all such characters into ASCII labels.

IDNs (Internationalised Domain Names)

IDNs are domain names represented by local language characters. Such domain names could contain characters with diacritical marks as required by many European languages, or characters from non-Latin scripts (for example, Arabic or Chinese).

IDNs made the domain name label as it is displayed and viewed by the end user different from that transmitted in the DNS. To avoid confusion the following terminology is used:

The A-label is what is transmitted in the DNS protocol and this is the ASCII-compatible (ACE) form of an IDNA string; for example, “xn--11b5bs1di”.

The U-label is what should be displayed to the user and is the representation of the IDN in Unicode; for example “□□□□□“ (“test” version in Hindi, Devanagari script).

Lastly, the LDH-label strictly refers to an all-ASCII label that obeys the “hostname” (LDH) conventions and that is not an IDN; for example, “icann” in the domain name “icann.org”.

(The above label definition is extracted from: http://www.ietf.org/internet-drafts/draft-klensin-idnabis-issues-01.txt)

If you are interested in learning more about the status of IDNs or how to participate in the work to deploy IDNs, please contact ICANN’s IDN Director, Tina Dam (tina.dam@icann.org), or participate in one of the IDN sessions this week.

A list of all IDN-related sessions in San Juan is available at: http://sanjuan2007.icann.org/event/2007/06/23/table/all/2

http://icann.org