GA2002/13-11

International Domain Names

Briefing draft document for discussion

Elisabeth Porteneuve, 19 February 2002

Introduction

The International Domain Names issue comes to lights less than two years ago, as the resultant of several actions:

1. Competition by gTLD Registries towards new customers (cf. ICANN Board resolutions from September 2000, http://www.icann.org/minutes/prelim-report-25sep00.htm).

2. Natural need of people from the worldwide Internet asking to facilitate their access to the Internet domain names.

3. General dissatisfaction of the worldwide Internet with ICANN and its incapability to became international body, triggering off a strong reactions from various horizon, including requests for international characters in domain names.

Subsequently the IETF has been asked to make a technical study on International Domain Names introduction into the DNS.

Background to International Domain Names

1. ASCII subset "LDH"

   Today domain names code specifications limit the permissible code points to a restricted subset of 38 signs: the letters a-z (upper and lower case alike, 26 signs), the digits 0-9, the hyphen-minus "-" (so called "LDH"), plus the label-separating period (with additional rules such as no minus at the beginning or at the end of a label).

2. Unicode

   The Unicode is the only existing, very recent, table of international character sets produced originally by printer's industry in late 1980's.

   The origins of Unicode are rooted in works on unified Han, a subset of Chinese, Japanese and Korean (CJK) characters, which (1) have identical internal computer code point, (2) print in Chinese, Japanese or Korean design, according to a language context, (3) may have a similar meaning or not, according to a language context.

   An important characteristic of three CJK languages is that they do not use a small set of signs called alphabet, but rather write ideographs, each of them being a concept or a word.

   There is more than one hundred thousands of CJK ideographs.

   The Unicode consortium spent ten years on developing unified Han for printer's industry.

   The Unicode tables allows for up to 4 octets per character.

   The Unicode Consortium tables are equivalent to ISO 10646 tables. The ISO Working Group responsible for ISO/IEC 10646 is JTC1/SC2/WG2.

3. The Traditional Chinese vs Simplified Chinese issue:

After the establishment of the People's Republic of China in 1949, and to ensure the coherency of a large country with the largest world population, the new regime led by Mao Zedong and Zhou Enlai announced that character simplification was a high priority task. As a result of works on Chinese simplification, various written language reforms were undertaken, the most important of which include: the development of a standardized translation to Latin system known as *pinyin* (cf. replacement of the old name Pekin by Beijing), and the simplification of thousands of ideographs forms.

Since the promulgation of simplification plan of Chinese characters in 1956, simplified Chinese has been used in Mainland of China; the plan has greatly promoted the spread and application of the Chinese language in people's daily life, meanwhile, the traditional Chinese still extensively exist in the social life, for its long history and artistic value.

After the reunification of Hong Kong to the Mainland of China in 1997 and Macao in 1999, both simplified and traditional Chinese characters are used widely in the these two regions. Singapore mainly uses simplified Chinese. People in Taiwan region have always used traditional Chinese characters. Korea and Japan mainly use traditional Chinese.

4. The IETF works on International Domain Names

   a. Based on Unicode, because there is nothing else.

   b. Technical scope - expand today "LDH" 38 characters into several tens of thousands of code points.

   c. Led to a discovery of many problems, two of them are listed below:

      - Combinatory effects which may have a dramatic impact on domain names. At the extreme stage both an end user and a business company being unable to communicate without knowing precisely which language code points were used to print business cards or publish web sites - in other words neither an end user nor a business company can use a printed information and enter it into browser without knowing which scripts must be used.

      - Mutual incompatibility in Unicode between unified Han (developed for printers) and Chinese language including Simplified Chinese.

Summary of problems and political dilemmas:

1. Chinese Unicode problem:

   a. It is mutually impossible to satisfy TC/SC _and_ unified Han (all engineers worldwide affirm it)

   b. Chinese are signatories to ISO10646 (ISO documents are signed by official representatives), therefore some presume, in all due deference to governments, that Chinese decision gives advantage to unified Han over TC/SC problem. But many forget that initially Unicode and unified Han were made for printer's industry, in 1980's, at the time when computer memory and processing were slow and requested for a lot of ingenuity to allow new features. Consequently Chinese could endorse ISO10646 for printers, and may be against its usage for domain names. On the other hand there is no doubt that Chinese endorse their national work on Chinese language Simplification, and the fact that one billion nation uses it now.

   c. The IETF work on Unicode usage for domain names demonstrates a clash between Chinese language on one side and Korean and Japanese on another side. In other words accepting Unicode for domain names is an equally bad choice between supporting Korean and Japanese against Chinese, or an opposite. None of these is either wise or appropriate.

2. Latin - Cyrillic - Greek problem:

a. If the usage of mixed letters from various alphabets is allowed - and the IETF works on Unicode characters cannot exclude it - then, there will be no more any unambiguous printed URL. The mixed similarly appearance while different code points will create a terrible confusion to consumers, and may kill any hope for safe electronic commerce.

b. The combinatory possibilities will increase by factor of hundred or thousand a domain names cost to some companies

c. The Latin - Cyrillic - Greek problem is one example among others. The combinatory effects are similar for CJK as well.

3. The UDRP for IDN problem:

a. The "language" for IDN in gTLD is undefined.

b. The printed URLs (paper or screen) are in general case undefined because a multitude of characters in different scripts have the same printed shape. As un example a business card with abc.com in IDNs does not provide for unilateral guessing of a company. In case of any trading services a consumer can be easily abused, and he will be not able to demonstrate so.

4. Not enough work on languages:

a. The Unicode is a recent, unique, pot pourri, initially defined for printer's industry, gathering not only languages, but anything which may be printed.

b. But there is nothing else. Nothing dedicated to languages for international domain names.

5. Tradeoffs

a. Do we consider that from a user perspective an alternate roots are more confusing that IDNs in gTLDs ?

b. If trademark industry or business or electronic commerce industry feels in danger with IDNs, odds are there will be pressure to create as many "LDH" TLDs as companies, and allow them to escape this way from combinatory effects and confusion.

Questions :

1. Do we agree the Unicode is unsuitable for Internet Domain Names on global Internet. ?

2. If yes, what to do now ?

Bibliography:

1. The Internationalized Domain Names IETF WG http://www.ietf.org/html.charters/idn-charter.html
The IDN IETF WG Web site http://www.i-d-n.net/
The IDN IETF mailing list archives ftp://ops.ietf.org/pub/lists/idn.current

2. IETF drafts related to IDN:
a. [AMC-ACE-M] Adam Costello (4 Sep 2001)
http://www.ietf.org/internet-drafts/draft-ietf-idn-amc-ace-z-01.txt
The choice of AMC-ACE encoding got a significant support within
Internet industry. Subsequently its name became PUNYCODE:
b. [PUNYCODE] Adam Costello (10 Jan 2002)
http://www.ietf.org/internet-drafts/draft-ietf-idn-punycode-00.txt
c. [IDNA] P. Faltstrom, "Internationalizing Domain Names in Applications" (11 Jan 2002)
http://www.ietf.org/internet-drafts/draft-ietf-idn-idna-06.txt
d. [NAMEPREP] Paul Hoffman and Marc Blanchet, "Stringprep Profile for Internationalized Host Names" (17 Jan 2002)
http://www.ietf.org/internet-drafts/draft-ietf-idn-nameprep-07.txt

  e. [TC/SC] XiaoDong Lee, Hsu Nai-Wen, Deng Xiang, Erin Chen, Zhang Hong, Sun Guonian, "Traditional and Simplified Chinese Conversion" (16 Nov 2001)
  http://www.ietf.org/internet-drafts/draft-ietf-idn-tsconv-02.txt

3. IETF RFCs:
   a. [RFC3066] H. Alvestrand, "Tags for Identification of Languages"
   http://www.ietf.org/rfc/rfc3066.txt

4. Unicode Consortium:
   [UNICODE] The Unicode Standard, Version 3.1.0: The Unicode Consortium.
   http://www.unicode.org/charts

5. The Pitfalls and Complexities of Chinese to Chinese Conversion, J.Halpern and J.Kerman,
   http://www.cjk.org/cjk/c2c/c2cbasis.htm

6. ICANN Board IDN Committee:
   http://www.icann.org/committees/idn/

7. Names Council IDN Task Force:
   http://www.dnso.org/clubpublic/nc-idn/Arc00/