Internet Draft Liana Ye
draft-ietf-idn-step-00.txt Y&D ISG
May 29, 2001
Expires in six months (November 2001)

StepCode- A User Access Oriented IDN Encoding

Status of this memo

Abstract

This document describes a transformation method from an end user into Unicode library for representing Chinese characters in host name parts in a fashion that is extendable to include more than Unicode symbols and completely compatible with the current DNS. It is a potential candidate for an ASCII-Compatible Encoding (ACE) for internationalized hostnames. This method is based on user widely used concept for denoting a Chinese character with its phonetic elements, and register their IDN names with such an extented phonetic description (English speakers can register a CJK glyph for their blessing too.), so that an IDN name in either traditional or simplified character set can be effective with both user communities depending on servers, and allows standard variations for compressing and security filtering of that information. The method can be extented for other scripts as long as they need more then 26 basic ASCII letters to be mapped.

1. Introduction

1.1 Context

There is a strong world-wide desire to use characters other than plain ASCII in hostnames. Hostnames have become the equivalent of business or product names for many services on the Internet, so there is a need to make them usable by people whose native scripts are not representable by ASCII. The requirements for internationalizing hostnames are described in the IDN WG's requirements document, [IDNReq].

The IDN WG's comparison document [IDNComp] describes three potential main architectures for IDN: arch-1 (just send binary), arch-2 (send binary or ACE), and arch-3 (just send ACE). StepCode is an ACE that can be used with protocols that match arch-2 or arch-3. Hopefully, this ACE may render arch-2 unnecessary. It is either a server or a client generatable and selectable ACE according to a string's language/script tag/label[RFC1766]. Because it does not attempt any particular optimization or compression of string patterns, the average length for a Chinese character in [ISO10646] is about 13 characters long, enough to code ÏÃÙ±Ì ©¬.com (I wish 20% reader can see this, and the rest should be at lost.) as: xinzhuqinghua1212qin1jin0ge1ge0shui1qing0hua2shi0.com with 54 octets. In the code the first group of digits specify the tones of each glyph as well as the number of codepoints in [ISO10646] or glyphs. The rest of digits specify the layout of different parts of each glyph of above string.

The StepCode protocol has the following features:

- There is exactly one way to convert internationalized host parts to and from StepCode strings with a script tag/label. It permits different script tags to access the same glyph in [ISO10646] similar with a searching method into a book library, while the codepoints of [ISO10646] is an analogy to ISDN book number. Host name part uniqueness is preserved. If there is a difference in the code, it is considered as user input error.

- Host parts have no international glyphs but US-ASCII.

- Names using StepCode have lengths proportionate to the number of glyphs (from IS 10646) in the names themselves plus the script tag. However, StepCode for most frequently used glyphs in the table is shortend significantly such as Cyrillic.

- This specification allows standard compression or security treatment compatible with existing hostnames.

It is important to note that the following sections contain many normative statements with "MUST" and "MUST NOT". Any implementation that does not follow these statements exactly is likely to cause damage to the Internet by creating non-unique representations of hostnames.

1.2 Author's Disclaimer

This document was written for the convenience of the IDN WG, in case someone believes that there are no agreeable mechanisms for referencing internationalized names without converting [ISO10646] codepoints. The author believes that [ISO10646] is to establish a symbol libaray, and there are better ways to do high frequency symbol accessing. Display a symbol onto no DNS-based approach can solve the "IDN" problem as it is hoped by users and company/enterprise domain name registrants and it is prossible not to add another coding format, further complicate the DNS and risk unknown problems and incompatibilities.

1.3 Terminology

The key words "MUST", "SHALL", "REQUIRED", "SHOULD", "RECOMMENDED", and "MAY" in this document are to be interpreted as described in [RFC2119].

Hexadecimal values are shown preceded with an "0x". For example, "0xa1b5" indicates two octets, 0xa1 followed by 0xb5. Binary values are shown preceded with an "0b". For example, a nine-bit value might be shown as "0b101101111".

Examples in this document use the notation from the Unicode Standard [Unicode3] as well as the ISO 10646 names. For example, the letter "a" may be represented as either "U+0061" or "LATIN SMALL LETTER A".

StepCode converts strings at a client site with internationalized characters into strings of US-ASCII that are acceptable as host name parts in current DNS host naming usage. The former are called "pre-converted" and a "glyph" for a symbol repesented by one codepoint in [ISO10646] or "glyphs" for a string of that, and the latter are called "post-converted".

The protocol contains one procedure and calls for standardizing a minimum number of glyphs of a script using the same script tag. Glyphs in the minimum number of glyph set is called "particles".

The protocol using US-ASCII to denote the phonetic elements of a script and calls for standardizing such a mapping for each script tag. The phonetic elements of a glyph is called "spelling" of the glyph and is called "stem" for that of a particle.

The protocol specifies an ASCII Compatible [ACE] Encoding method and using Chinese script as an example to demonstrate its features, here is refered as an "ACE" process.

1.4 IDN summary

Using the terminology in [IDNComp], StepCode specifies an ACE format for arch-2 (send binary or ACE), and arch-3 (just send ACE).

The length characteristics of StepCode are discussed above (1.1 Context), is a variable depending on users' choice among many factors. It fits well with existing compression and security treatments.

It calls for standardrizing phonetic elements and minimum glyph set within its user community (and labeled by script tag), while asking the internet industry to enforce the standard and providing cross reference to different script tags into Unicode standard.

2. Host Part Transformation

According to [STD13], host parts must be case-insensitive, start and end with a letter or digit, and contain only letters, digits, and the hyphen character ("-"). This, of course, excludes any internationalized characters, as well as many other characters in the ASCII character repertoire. Further, domain name parts must be **63 octets or shorter in length.**

2.1 Name tagging

All post-converted name parts that contain internationalized characters begin with the string "gl-p-", where "gl" denote the glyph set or script encoded as specified by [RFC2277], and "p" denote the phonetic standard used, it SHOULD be reserved with IANA. The string "gl-p-" will allow 674 scripts and 24 phonetic standards of each to be encoded. The assignment of "gl-p-" shall be defined in future versions of this draft.

Note that a zone administrator MAY still choose to use "gl-p-" at the beginning of a hostname part even if that part does not contain internationalized characters. Zone administrators MAY create host part names that begin with "gl-p-" which means no conversion is done and display systems SHOULD ignore converting internationalized characters back for display.

2.2 Converting an internationalized name to an ACE name part

To convert a string of internationalized characters into an ACE name part, the following steps MUST be preformed in the exact order of the subsections given here.

If a name part consists exclusively of characters that conform to the hostname requirements in [STD13] or the string "gl-p-", the name MUST NOT be converted to StepCode. That is, a name part that can be represented without StepCode MUST NOT be changed. This absolute requirement prevents:

1. double encoding from a client of user keyboard input and a server provider;

2. mess up existing registered domain names;
3. from being two different encodings for a single DNS registered hostname;
4. interfering with registered glyphs with more than one phonetic standard, such as Chinese script.

If any checking for prohibited name parts (such as ones that are prohibited characters, case-folding, or canonicalization) is to be done, it MUST be done after doing the conversion to an ACE name part as it is specified in [nameprep].

Characters outside the first plane of characters (those with codepoints above U+FFFF) MUST be represented using surrogates, as described in the UTF-16 description in ISO 10646.

The input name string consists of characters from the ISO 10646 character set in big-endian UTF-16 encoding. This is the pre-converted string.

2.2.1 Check the input string for disallowed names

If the input string consists only of characters that conform to the hostname requirements in [STD13], or the input string consists a null language tag, the conversion MUST stop with an error.

2.2.2 Represent glyphs by their spelling and particle layout.

2.2.2.1 StepCode defination for digits

Tone marks [Macmillan93]
0 no tone
1 flat/macron (-)
2 rise/acute (/)
3 dip/breve (v)
4 drop/grave (\)
5 throw/circumflex (^)
6 thrill/tilde (~)
7 dieresis (..)
8 cedilla (hook)

Particle layout [Ye95]
0 end of a stem or a spelling
1 to its left
2 to its bellow
3 to its inside (an enclosure particle)
4 to its outside (normally a center divider)

2.2.2.2 StepCode phonetic symbol tables

2.2.2.2.1 Chinese

Note: This is a list extracted from [Dictionary79] and

other sources.

Pinyin	Wade	Zhuyin
b	p	b
p	p/	p
m	m	m
f	f	f
d	t	d
t	t/	t
n	n	n
l	l	l
g	k	g
k	k/	k
h	h	h
j	ch	j
q	ch/	q
x	hs	x
zh	ch	zh
ch	ch/	ch
sh	sh	sh
r	j	r
z	ts	z
c	ts/	c
s	s	s
a	a	a
o	u	o
e	e^	e
e^	eh	ei
i	uh	i
u	u	u
uo	o	oo
u..	u..	u..
y	y	y
w	w	w

v (' spelling separator)

2.2.2.2.2 Other scripts to come

2.2.3 StepCode Format

Format Defination: A Stepcode unit is a string of [A-Za-z0-9] characters without any white spaces, BLANK, in between. For each StepCode unit, there are data elements indicated by "", which is a required elelment, and [] where the element is optional, and / where the data is selectable.

Sx stands for Spelling of xth glyph;
Tx stands for Tone of xth glyph;
Py stands for Stem for yth particle;
Ly stands for Layout relation from y to y+1;
Px.y stands for Stem for Xth glyph and its yth particle;
Lx.y stands for Layout relation from Xth glyph and its y to y+1.

2.2.3.1. One glyph

"S""T"[P1][L1][P2][L2]...[Py][0/BLANK]

Example:xin1qin
xin1qin1jin0

2.2.3.2. Glyphs

"S1S2S3...Sx""T1T2...Tx"[P1.1][L1.1][P1.2][L1.2]...[P1.y][0]
[P2.1][L2.1][P2.2][L2.2]...[P2.y][0]
...
[Px.1][Lx.1][Px.2][Lx.2]...[Px.y][0/BLANK]

Example of glyphs of four:
xinzhuqinghua1212
xinzhuqinghua1212qin1jin0ge1ge0shui1qing0
xinzhuqinghua1212qin1jin0ge1ge0shui1qing0hua
xinzhuqinghua1212qin1jin0ge1ge0shui1qing0hua2shi0

The four StepCodes are equivalent, depending on where
it is registered, the size of the database, as well as there exist
similar hostnames it has confict with.

2.3. StepCode Encoding Process

Either, StepCode may be obtained from Unicode to StepCode through
a code lookup table, and combines glyph code into glyphs code as
shown in 2.2.

Or, it is inputed directly from keyboards, where an input
processing module to verify correctness of intented glyphs is necessary.

Prepend "glp--" or the name of conversion table used as script tag
to the post-convered string; finish. This is the hostname part that
can be used in DNS registration as well as resolution.

Go through [nameprep], checking for prohibited characters,
case-folding, or canonicalization.

2.4. Converting a StepCode hostname part to an internationalized name

The process has three steps with script tag untouched:

Step 1.If a domain name part consists no script tag, then goto Step 3;
Otherwise enable conversion table named "glp" from StepCode to
Unicode or other code, obtain the corespondends.
Step 2.If the corespondent is there then goto Step 3;
Otherwise decomposes the post-converted string into a number
of individual glyph specified in the "T" field;
Searching for each glyph;
If any of the glyph is not found,
compose an error message and
Requesting the missing glyph to be supplied

from the sender.
Step 3.Display available glyph where missing glyph is shown with its StepCode.

3. Security Considerations

Much of the security of the Internet relies on the DNS. Thus, any change to the characteristics of the DNS can change the security of much of the Internet. Thus, StepCode makes no changes to the DNS itself.

Hostnames are used by users to connect to Internet servers. The security of the Internet would be compromised if a user entering a single internationalized name could be connected to different servers based on different interpretations of the internationalized hostname. Thus the restriction of DNS names to a small symbol set is necessary and effective, where adding any other data format such as UTF-8 only opens the security gate for complication.

4.Internationalization considerations

StepCode is designed so that every internationalized hostname part can be represented as one and only one DNS-compatible string. If there is two different ways to obtain the same glyph on a display device, then they are still two distinct hostnames, with no bearing on security issue. If there is any way to follow the steps in this document and get two or more different results, it is an error in the domain name registration process, where one domain name register fails updates other domain name register servers a newly registered and well researched hostname.

5. References

[ASCII] American National Standards Institute (formerly United States of America Standards Institute), X3.4, 1968, "USA Code for Information Interchange". (ANSI X3.4-1968)

[IDNCOMP] "Comparison of Internationalized Domain Name Proposals", draft-ietf-idn-compare-00.txt, June 2000, P. Hoffman.

[IDNReq] Zita Wenzel and James Seng, "Requirements of Internationalized Domain Names", draft-ietf-idn-requirements. May 2001.)

[ISO10646] ISO/IEC 10646-1:2000 (note that an amendment 1 is in preparation), ISO/IEC 10646-2 (in preparation), plus corrigenda and amendments to these standards.

[Dictionary79] Beijing Foriegn Language Dept., "A Chinese-English Dictionary", 1979, BK# 9017.810.

[Macmillan93] The Macmillan Visual Desk Reference, 1993, ISBN 0-02-531310-x.

[RFC2277] "IETF Policy on Character Sets and Languages", rfc2277.txt, January 1998, H. Alvestrand.

[RFC2119] Scott Bradner, "Key words for use in RFCs to Indicate Requirement Levels", March 1997, RFC 2119.

[STD13] Paul Mockapetris, "Domain names - implementation and specification", November 1987, STD 13 (RFC 1035).

[UNICODE] The Unicode Consortium, "The Unicode Standard". Described at http://www.unicode.org/unicode/standard/versions/.

[UNICODE30] The Unicode Consortium, "The Unicode Standard -- Version 3.0", ISBN 0-201-61633-5. Same repertoire as ISO/IEC 10646-1:2000. Described at http://www.unicode.org/unicode/standard/versions/Unicode3.0.html.

[Ye95] Liana Ye, "A Language Oriented Chinese Encoding for Multilingual Computing Environments", in "Proceeding of the 1995 International Conference on Computer Processing of Oriental Languages", Page 323.

6. Acknowledgements

The author has reused existing IDN draft document and language as much as possible to demonstrate the deep respect for the work has been done by members of this working group.

7. IANA Considerations

This document require IANA action for availibility of language tag, and registration for each tag and its sub-field for phonetic system used.

8. Author Contact Information

Liana Ye
Y&D ISG
2607 Read Ave.
Belmont, CA 94002, USA.
(650) 592-7092
liana.ydisg@juno.com