INTERNET-DRAFT
draft-ietf-idn-hangeulchar-00.txt
Expires 2001-Dec-27

Soobok Lee
GyeongSeog Gim

2001-Jun-27

Hangeul NAMEPREP recommendation version 1.0

Status of this Memo

Abstract

This document suggests Hangeul-specific NAMEPREP recommendations.
It defines :

- mapping tables for half-width jamo and enclosed jamo
- excluding compatibility Hangeul jamo block
  from compatibility decomposition in normalization step
- criteria for determining invalid syl-ipf jamo sequence
- prohibited hangul filler character in KC norm output.

Contents

Overview

A user can enter a domain name into an application program in a myriad of fashions and the characters entered in the domain name may or may not be those that are allowed in internationalized host names. Thus, there must be a way to normalize the user's input before the name is resolved in the DNS, which is the rationale for NAMEPREP.

NAMEPREP design goals are :

- to allow users to enter host names in applications and have the highest chance of getting the name correct. The user should not be limited to only entering exactly the characters that might have been used for domain name registration, but be able to enter characters that can be unambiguously normalized to characters in the registered domain name.

- to prohibit as few characters as possible that might be used in the future and in the various contexts

- to allow the widest possible set of host names as long as those host names do not cause other problems, such as conflict with other standards.

The NAMEPREP process to prepare internationalized host names for use in the DNS includes the following stages :

- stage1 : mapping characters to other characters, such as to change their case, mapping out some meaningless characters

- stage2 : normalizing characters using normalization form KC. KC form consists of two steps detailed in [UTR15]
  - compatibility decomposition
  - canonical composition

- stage3 : excluding characters that are prohibited from appearing in internationalized host names

This draft defines special Hangeul character mappings and exceptions in applying KC normalization. And this draft also defines some prohibited Hangeul characters and sequences so that Hangeul can be used safely in Internet identifiers such as IDN.

The content of this draft is subject to change with further discussions and studies.

Background : UCS Hangeul

Korean Hangeul syllables are formed from a set of Hangeul letters, called jamo in Korean, in a regular fashion.

The ISO/IEC 10646 (=Unicode Standard) contains both the complete set of precomposed modern Hangeul syllable blocks and the set of syl-ipf Hangeul jamo (= conjoining jamo in [UNICODE]). This set of syl-ipf jamo can be used to encode all modern and old syllable blocks. For a description of syl-ipf Hangeul jamo behavior and precomposed Hangeul Syllables, see [UNICODE].

Hangeul jamo are divided into three classes: choseong (leading consonants), jungseong(vowels), and jongseong(trailing consonants). In the following paragraphs, these classes are abbreviated as L (leading consonant), V(vowel), and T (trailing consonant). And for use in composition, two invisible filler characters act as placeholders for choseong or jungseong:
U+115f (Hangeul choseong filler) and
U+1160 (Hangeul jungseong filler).

The UCS/Unicode contains a set of Hangeul Compatibility jamo (U+3130~U+318F) which consists of a filler, nonsyl-ipf Hangeul consonants and vowel elements. These characters are provided solely for compatibility with the KS X 1001 (formerly KS C 5601) standard. Unlike the characters found in the Hangeul jamo block (U+1100 .. U+11FF), the compatibility jamo characters have no syl-ipf semantics, except for only their filler+L+V+T or filler sequence makes a Hangeul syllable according to KS X 1001.

The UCS/Unicode Standard also contains 52 half-width modern Hangeul jamo in the halfwidth and fullwidth forms (U+FFA0 .. U+FFDC) block and enclosed Hangeul syllables and jamo in the enclosed CJK letter and month block (U+3200 .. U+32FF). Enclosed ones are consisted of parenthesized jamo and circled jamo.

Hangeul canonical composition

Modern Hangeul syllables can be expressed with either two or three jamo, either in the form consonant + vowel or in the form consonant + vowel + consonant. There are 19 possible leading (initial) consonants (choseong), 21 vowels (jungseong), and 27 trailing (final) consonants (jongseong). Thus there are 399 possible two-jamo syllables and 10,773 possible three-jamo syllables, for a total of 11,172 modern Hangeul syllables.

Each of the Hangeul syllables may be encoded by an equivalent

sequence of syl-ipf jamo; however, the converse is not true because thousands of archaic Hangeul syllables may be encoded only as a sequence of syl-ipf jamo. Implementaions that use a syl-ipf jamo encoding are able to represent these archaic Hangeul syllables.

The Hangeul syllables can be derived from syl-ipf jamo by a regular process of composition. The algorithm that maps a sequence of syl-ipf jamo to the encoding point for a Hangeul syllable is detailed in [UNICODE].

In canonical composition, the syl-ipf jamo sequence for modern Hangeul syllable is transformed into the modern Hangeul syllable, but the sequence for archaic Hangeul syllable and the invalid jamo sequence (defective combining character sequence) are preserved in this process.

In normalization form KC, all input sequence of code points go through this canonical composition [UTR15]. If any invalid jamo sequence is detected after KC normaliation stage, as it is not displayable correctly and distinguishably, the sequence should be prohibited from being an identifier. Whether a syl-ipf jamo sequence is valid or not can be determined according to the criteria detailed in [UNICODE].

Hangeul compatibility decomposition

In normalization form KC, all input code sequence go through this compatibility decomposition and then canonical composition.

Every Hangeul compatibility jamo and half-width jamo have its corresponding compatibility equivalent Hangeul syl-ipf jamo defined in [UNICODE_CHART].

But this equivalence does violate the semantics and combining rules for compatibility jamo sequence in [KSC5601] from which UCS compatibility jamo came.

In [KSC5601], a valid compatibility jamo sequence should start with a filler followed by choseong,jungseong and jongseong (or filler) to denote a Hangeul syllable. If the sequence does not fulfill this criterion, its jamo should remain unchaged as compatibility jamo. The same for half-width Hangeul jamo.

Current compatibility decomposition blindly transforms compatibility jamo sequence even without a leading filler on a jamo by jamo basis. For example, a valid jamo sequence "filler gi-eog a gi-eug" (U+3164 U+3131 U+314F U+3131) denoting a Hangeul syllable "gag"(U+AC01) is errornously transformed into "jungsong_filler chosung_gi-eog jungseong_a chosung_gi-eog" (U+1160 U+1100 U+1161 U+1100) that are canonically composed into "syllable_ga choseong_gi-eog"

(U+AC00 U+1100) which are false.

If false composistion could be avoided, NAMEPREP should exclude compatibility jamo and half-width jamo from its compatibility decomposition step. And, only valid compatibility jamo sequence should be recognized and transformed into a syl-ipf jamo sequence at the mapping step before KC normalization step in NAMEPREP.

Hangeul consonant sequence can be used as abbreviated form of long Hangeul syllables sequence that represent Hangeul business name. And, there may be future need to represent Hangeul syllables in compatibility jamo sequences for an alternative syllable writing/ displaying scheme.

In NAMEPREP KC normalization and its internal compatibility decomposition step, each parethesized Hangeul jamo is transformed into its compatibility equivalent character sequence consisted of one pair of parentheses with inner Hangeul jamo and then that sequence is treated as an invalid domain since the paranthesis is prohibited in the domain names. For example, parenthesized gi-eog (U+3200) is decomposed into U+0028 + U+1100 + U+0029 which includes prohibited left and right parentheses (U+0028,U+0029 respectively).

Each parethesized Hangeul syllable is transformed into its compatibility equivalent character sequence consisted of one pair of parentheses with inner Hangeul syllable and then that sequence is treated as an invalid domain since the paranthesis is prohibited in the domain names.

In NAMEPREP KC normalization and its internal compatibility decomposition step, Circled Hangeul jamo is transformed into its compatibility equivalent Hangeul jamo which is not appropriate in IDN context, and preferrably, this NAMEPREP process should map this circled one into the corresponding compatibility Hangeul jamo before KC normalization to bypass this inappropriate compatibility decomposition.

Summarized Recommendations

KC normalization employed in NAMEPREP process does not preserve some Hangeul code semantics and so we recommend the following additional NAMEPREP actions for Hangeul codes:

* Stage 1: mapping

  - circled Hangeul jamo
    = map into the corresponding Hangeul compatibility jamo
      code range: U+3160 ~ U+326D
      mapping table detailed in appendix 3.

```
    -  half-width Hangeul jamo
       = map into the corresponding Hangeul compatibility jamo
         code range: U+FFA0 ~ U+FFDC
         mapping table detailed in appendix 4.

    -  transform compatibility jamo sequence with leading filler
       (U+3164) into syl-ipf jamo sequence
       = if and only if
         the sequence is of filler+ L+ V+ T (or filler) form.
       = preserve unchanged if the sequence is not of this form
       = so that each resulting jamo is given intended choseong
         or jongseong semantics implied in the input sequence

  * Stage 2: KC normalization

    -  compatibility decomposition
       = exclude compatibility Hangeul jamo; preserve them
         code range: U+3130 ~ U+318F

  * Stage 3: prohibitions

    -  prohibit invalid syl-ipf Hangeul jamo sequences
       = return error if not meaningful LV or LVT sequence

    -  compatibility Hangeul filler (U+3164) not combined
       = return error


Comments on security implication of inter-lingual similarities

  We have found  many similarities between hangeul jamo and
  other language scripts like japanese katakana and latin.

  To list some of them:

   - hangeul jamo gi-eog  and katakana hu
   - hangeul jamo mi-eum  and katakana ro
   - hangeul jamo i-eung  and latin 'o'
   - hangeul jamo ji-euth and katakana su
   - hangeul jamo ki-eog  and katakana wo
   - hangeul jamo a       and katakana to

   - hangeul syllable ma  and katakana ro-to
   - hangeul syllable ja  and katakana su-to
   - hangeul syllable ga  and katakana hu-to
   - hangeul syllable i   and digits '01'

  Some hangeul domains similiar to katakana domains
  can mislead some japanese to believe hangeul hostnames or
  hangeul email addresses are the japanese ones they trust.

  To mitigate these inherent security problems, there should be
```

well-prepared registration/dispute resolution policy that
can be enforced to every zone masters (including root zone
and its lower-level zones) and every email account masters.
Of course, whether this is feasible or not is beyond NAMEPREP scope.

Security considerations

This suggestion improves IDN security by prohibiting/correcting
non-displayable or invalid hangeul syllables/sequences in IDN.

References

[IDNREQ] Requirements of Internationalized Domain Names
http://www.ietf.org/internet-drafts/draft-ietf-idn-requirements-08
.txt

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

[UNICODE_CHART] THe Unicode Code Charts
http://www.unicode.org/charts/

[IDNA] Patrik Falstrom, Paul Hoffman,
"Internationalizing Host Names In Applications (IDNA)",
http://www.ietf.org/internet-drafts/draft-ietf-idn-idna-02.txt

[NAMEPREP] Paul Hoffman, Marc Blanchet,
"Preparation of Internationalized Host Names", Feb 2001,
http://www.ietf.org/internet-drafts/draft-ietf-idn-nameprep-03.txt

[UTR15] Mark Davis and Martin Duerst.
Unicode Normalization Forms. Unicode Technical Report;15.
http://www.unicode.org/unicode/reports/tr15/

[VERSION] M Blanchet
"Handling versions of internationalized domain names protocols",
http://www.ietf.org/internet-drafts/draft-ietf-idn-version-00.txt

[ISO10646] ISO/IEC, Information Technology - Universal
Multiple-Octet Coded Character Set (UCS) - Part 1: Architecture
and Basic Multilingual Plane, Oct. 2000, with amendments.

[KSC5601] Korean Standard KS C 5601- 1987

A1. Acknowledgements

Dongman Lee <dlee@icu.ac.kr> and Yangwoo Ko <newcat@peacenet.or.kr>
made valuable contributions to narrowing down the issues of the
prohibition and preservation of some hangeul characters.

Thank Mark Davis for his advice on useful UNICODE reference documents.

A2. Authors

Soobok Lee <lsb@postel.co.kr>
Postel Services, Inc.
http://www.postel.co.kr
Tel: +82-11-9774-2737

GyeongSeog Gim <gimgs@asadal.pusan.ac.kr>
Department of Computer Engineering
Pusan National University
Republic of Korea
Tel: +82-51-510-2292

A3. the mapping table for enclosed jamo in the format of [VERSION]

```
version=1.0

3260;1.0;3131
3261;1.0;3134
3262;1.0;3137
3263;1.0;3139
3264;1.0;3141
3265;1.0;3142
3266;1.0;3145
3267;1.0;3147
3268;1.0;3148
3269;1.0;314A
326A;1.0;314B
326B;1.0;314C
326C;1.0;314D
326D;1.0;314E
```

A4. the mapping table for half-width jamo in the format of [VERSION]

```
version=1.0

FFA0;1.0;3164
FFA1;1.0;3131
FFA2;1.0;3132
FFA3;1.0;3133
FFA4;1.0;3134
FFA5;1.0;3135
FFA6;1.0;3136
FFA7;1.0;3137
FFA8;1.0;3138
FFA9;1.0;3139
FFAA;1.0;313A
```

FFAB;1.0;313B
FFAC;1.0;313C
FFAD;1.0;313D
FFAE;1.0;313E
FFAF;1.0;313F
FFB0;1.0;3140
FFB1;1.0;3141
FFB2;1.0;3142
FFB3;1.0;3143
FFB4;1.0;3144
FFB5;1.0;3145
FFB6;1.0;3146
FFB7;1.0;3147
FFB8;1.0;3148
FFB9;1.0;3149
FFBA;1.0;314A
FFBB;1.0;314B
FFBC;1.0;314C
FFBD;1.0;314D
FFBE;1.0;314E
FFC2;1.0;314F
FFC3;1.0;3150
FFC4;1.0;3151
FFC5;1.0;3152
FFC6;1.0;3153
FFC7;1.0;3154
FFCA;1.0;3155
FFCB;1.0;3156
FFCC;1.0;3157
FFCD;1.0;3158
FFCE;1.0;3159
FFCF;1.0;315A
FFD2;1.0;315B
FFD3;1.0;315C
FFD4;1.0;315D
FFD5;1.0;315E
FFD6;1.0;315F
FFD7;1.0;3160
FFDA;1.0;3161
FFDB;1.0;3162
FFDC;1.0;3163