Internet Draft
draft-ietf-idn-nameprep-00.txt
July 3, 2000
Expires in six months

Paul Hoffman
IMC & VPNC
Marc Blanchet
ViaGenie

Preparation of Internationalized Host Names

Status of this memo

Abstract

This document describes how to prepare internationalized host names for transmission on the wire. The steps include excluding characters that are prohibited from appearing in internationalized host names, changing all characters that have case properties to be lowercase, and normalizing the characters. Further, this document lists the prohibited characters.

1. Introduction

When expanding today's DNS to include internationalized host names, those new names will be handled in many parts of the DNS. The IDN Working Group's requirements document [IDNReq] describes a framework for domain name handling as well as requirements for the new names. The IDN Working Group's comparison document [IDNComp] gives a framework for how various parts of the IDN solution work together.

A user can enter a domain name into an application program in a myriad of fashions. Depending on the input method, 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 canonicalized

the user's input before the name is resolved in the DNS.

It is a design goal of this document to allow users to enter host names in applications and have the highest chance of getting the name correct. This means that the user should not be limited to only entering exactly the characters that might have been used, but to instead be able to enter characters that unambiguously canonicalize to characters in the desired host name. At the same time, this process must not introduce any chance that two host names could be represented by two distinct strings of characters that look identical to typical users. It is also a design goal to have all preprocessing of IDN done before going on the wire, so that no transformation is done in the DNS server space.

This document describes the steps needed to convert a name part from one that is entered by the user to one that can be used in the DNS.

1.1 Terminology

The key words "MUST", "SHALL", "REQUIRED", "SHOULD", "RECOMMENDED", and "MAY" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Examples in this document use the notation from the Unicode Standard [Unicode3] as well as the ISO 10646 [ISO10646] names. For example, the letter "a" may be represented as either "U+0061" or "LATIN SMALL LETTER A". In the lists of prohibited characters, the "U+" is left off to make the lists easier to read.

1.2 IDN summary

Using the terminology in [IDNComp], this document specifies all of the prohibited characters and the canonicalization for an IDN solution. Specifically, it covers the following sections from [IDNComp]:

prohib-1: Identical and near-identical characters
prohib-2: Separators
prohib-3: Non-displaying and non-spacing characters
prohib-4: Private use characters
prohib-5: Punctuation
prohib-6: Symbols
canon-1.2: Normalization Form KC
canon-2.1: Case folding in ASCII
canon-2.2: Case folding in non-ASCII

Note that this document does not cover:
canon-1.1: Normalization Form C
canon-2.3: Han folding

1.3 Open issues

This is the first draft of this document. Although there has been much discussion on the WG mailing list about the topics here, there has not

yet been much agreement on some issues. Now that there is a document to talk about, that discussion can be more focussed.

1.3.1 Where to do name preparation

Section 2.1 says to do name preparation in the resolver. An argument can be made for doing name preparation in the application, before the application service interface. An advantage of that proposal is that resolvers would not need to do any name preparation. A disadvantage is that applications would have to be updated each time the IDN protocol is updated, such as if new characters are added to the repertoire of allowed characters. It seems likely that resolvers are more easily updated than all the individual applications that use internationalized host names.

1.3.2 Choosing between normalization form C and KC

Much of the discussion of normalization on the WG mailing list assumed that normalization form C would be used. Near the time that this document was written, people started considering form KC instead of C. This document used form KC, but the reasons for doing so could be contentious.

1.3.3 Does the prohibition catch all bad characters?

On the mailing list, it was discussed doing prohibition in two steps: a short list of prohibited characters before case folding in order to prevent uppercase characters that have no lowercase equivalents from getting through, and then a full check on the output of normalization. In this draft, all checking is done before case folding, based on the (possibly wrong) assumption that none of the prohibited characters will re-appear after the case folding and normalization. If that assumption turns out to be wrong, a check for just those problematic characters can be added after normalization, or a full check against the prohibited characters can be added.

2. Preparation Overview

This section describes where name preparation happens and the steps that name preparation software must take.

2.1 Where name preparation happens

Part of the chart in section 1.4 of [IDNReq] looks like this:

```
+---------------+
| Application   |
+---------------+
      |  Application service interface
      |  For ex. GethostbyXXXX interface
+---------------+
```

In this specification, the name preparation is done in the resolver, before the DNS service interface. That is, it is acceptable for software in the application service interface (such as a "GetHostByName" API) to pass the resolver a name that has not been prepared. However, the resolver MUST prepare the name as described in this specification before passing it to the DNS service interface.

2.2 Name preparation steps

The steps for preparing names are:

1) Input from the application service interface -- This can be done in many ways and is not specified in this document

2) Look for prohibited input -- Check for any characters that are not allowed in the input. If any are found, return an error to the application service interface. This step is necessary to prevent errors in the following two steps. This step fulfills prohib-1, prohib-2, prohib-3, prohib-4, prohib-5, and prohib-6 from [IDNComp].

3) Fold case -- Change all uppercase characters into lowercase characters. Design note: this step could just as easily have been "change all lowercase characters into uppercase characters". However, the upper-to-lower folding was chosen because most users of the Internet today enter host names in lowercase. This step fulfills canon-2.1 and canon-2.2 from [IDNComp].

4) Canonicalize -- Normalize the characters. This step fulfils canon-1.2 from [IDNComp].

5) Resolution of the prepared name -- This must be specified in a different IDN document.

The above steps MUST be performed in the order given in order to comply with this specification.

3. Prohibited Input

Before the text can be processed, it must be checked for prohibited characters. There is a variety of prohibited characters, as described in this section.

Note that one of the goals of IDN is to allow the widest possible set of host names as long as those host names do not cause other problems, such as possible ambiguity. Specifically, experience with current DNS names have shown that there is a desire for host names that include personal

names, company names, and spoken phrases. A goal of this section is to prohibit as few characters that might be used in these contexts as possible while making sure that characters that might easily cause confusion or ambiguity are prohibited.

Note that every character listed in this section MUST NOT be transmitted on the DNS service interface. Although the checking is being performed before case folding and canonicalization, those steps cannot result in any of these characters if these characters are not in the input stream. [[[NOTE: THIS STATEMENT NEEDS TO BE CHECKED ALGORITHMICALLY.]]] If a DNS server receives a request containing a prohibited character, then the IDN protocol MUST return an error message.

Note that some characters listed in one section would also appear in other sections. Each character is only listed once.

3.1 prohib-1: Identical and near-identical characters

Many characters in [ISO10646] are identical or nearly identical to other characters. These were often included for compatibility with other character sets.

The characters prohibited because they are identical or nearly identical to allowed characters are:

00AD SOFT HYPHEN
00D7 MULTIPLICATION SIGN
01C3 LATIN LETTER RETROFLEX CLICK
02B0-02FF [SPACING MODIFIER LETTERS]
066D ARABIC FIVE POINTED STAR
1806 MONGOLIAN TODO SOFT HYPHEN
2010 HYPHEN
2011 NON-BREAKING HYPHEN
2012 FIGURE DASH
2013 EN DASH
2014 EM DASH
2160-217F [ROMAN NUMERALS]
FB1D-FB4F [HEBREW PRESENTATION FORMS]
FB50-FDFF [ARABIC PRESENTATION FORMS A]
FE20-FE2F [COMBINING HALF MARKS]
FE30-FE4F [CJK COMPATIBILITY FORMS]
FE50-FE6F [SMALL FORM VARIANTS]
FE70-FEFC [ARABIC PRESENTATION FORMS B]
FF00-FFEF [HALFWIDTH AND FULLWIDTH FORMS]

3.2 prohib-2: Separators

Horizontal and vertical spacing characters would make it unclear where a host name begins and ends. The prohibited spacing characters are:

0020 SPACE

00A0 NO-BREAK SPACE
1680 OGHAM SPACE MARK
2000-200B [SPACES]
2028 LINE SEPARATOR
2029 PARAGRAPH SEPARATOR
202F NARROW NO-BREAK SPACE
3000 IDEOGRAPHIC SPACE

Allowing periods and period-like characters as characters within a name part would also cause similar confusion. The prohibited periods, characters that look like periods, and characters that canonicalize to a period or to a period-like character are:

002E FULL STOP
06D4 ARABIC FULL STOP
2024 ONE DOT LEADER
2025 TWO DOT LEADER
2026 HORIZONTAL ELLIPSIS
2488 DIGIT ONE FULL STOP
2489 DIGIT TWO FULL STOP
248A DIGIT THREE FULL STOP
248B DIGIT FOUR FULL STOP
248C DIGIT FIVE FULL STOP
248D DIGIT SIX FULL STOP
248E DIGIT SEVEN FULL STOP
248F DIGIT EIGHT FULL STOP
2490 DIGIT NINE FULL STOP
2491 NUMBER TEN FULL STOP
2492 NUMBER ELEVEN FULL STOP
2493 NUMBER TWELVE FULL STOP
2494 NUMBER THIRTEEN FULL STOP
2495 NUMBER FOURTEEN FULL STOP
2496 NUMBER FIFTEEN FULL STOP
2497 NUMBER SIXTEEN FULL STOP
2498 NUMBER SEVENTEEN FULL STOP
2499 NUMBER EIGHTEEN FULL STOP
249A NUMBER NINETEEN FULL STOP
249B NUMBER TWENTY FULL STOP
33C2 SQUARE AM
33C2 SQUARE AM
33C7 SQUARE CO
33D8 SQUARE PM
33D8 SQUARE PM

3.3 prohib-3: Non-displaying and non-spacing characters

There are many characters that cannot be seen in the ISO 10646 character set. These include control characters, non-breaking spaces, formatting characters, and tagging characters. These characters would certainly cause confusion if allowed in host names.

0000-001F [CONTROL CHARACTERS]
007F DELETE
0080-009F [CONTROL CHARACTERS]
070F SYRIAC ABBREVIATION MARK
180B MONGOLIAN FREE VARIATION SELECTOR ONE
180C MONGOLIAN FREE VARIATION SELECTOR TWO
180D MONGOLIAN FREE VARIATION SELECTOR THREE
180E MONGOLIAN VOWEL SEPARATOR
200C ZERO WIDTH NON-JOINER
200D ZERO WIDTH JOINER
200E LEFT-TO-RIGHT MARK
200F RIGHT-TO-LEFT MARK
202A LEFT-TO-RIGHT EMBEDDING
202B RIGHT-TO-LEFT EMBEDDING
202C POP DIRECTIONAL FORMATTING
202D LEFT-TO-RIGHT OVERRIDE
202E RIGHT-TO-LEFT OVERRIDE
206A INHIBIT SYMMETRIC SWAPPING
206B ACTIVATE SYMMETRIC SWAPPING
206C INHIBIT ARABIC FORM SHAPING
206D ACTIVATE ARABIC FORM SHAPING
206E NATIONAL DIGIT SHAPES
206F NOMINAL DIGIT SHAPES
FEFF ZERO WIDTH NO-BREAK SPACE
FFF9 INTERLINEAR ANNOTATION ANCHOR
FFFA INTERLINEAR ANNOTATION SEPARATOR
FFFB INTERLINEAR ANNOTATION TERMINATOR
FFFC OBJECT REPLACEMENT CHARACTER
FFFD REPLACEMENT CHARACTER

3.4 prohib-4: Private use characters

Because private-use characters do not have defined meanings, they are prohibited. The private-use characters are:

E000-F8FF [PRIVATE USE, PLANE 0]

3.5 prohib-5: Punctuation

The following characters are reserved or delimiters in URLs [RFC2396] and [RFC2732]:

" # $ % & + , . / : ; < = > ? @ [ ]

3.5.1 Characters from URLs

The following punctuation characters are prohibited because they are reserved or delimiters in URLs.

0022 QUOTATION MARK
0023 NUMBER SIGN
0024 DOLLAR SIGN

0025 **PERCENT SIGN**
0026 **AMPERSAND**
002B PLUS SIGN
002C COMMA
002E FULL STOP
002F SOLIDUS
003A COLON
003B SEMICOLON
003C LESS-THAN SIGN
003D EQUALS SIGN
003E GREATER-THAN SIGN
003F QUESTION MARK
0040 **COMMERCIAL AT**
005B LEFT SQUARE BRACKET
005D RIGHT SQUARE BRACKET

3.5.2 Characters that canonicalize to characters from URLs

The following punctuation characters are prohibited because their normalization contains one or more of the characters from section 3.5.1.

037E GREEK QUESTION MARK
2048 **QUESTION EXCLAMATION MARK**
2049 **EXCLAMATION QUESTION MARK**
207A SUPERSCRIPT PLUS SIGN
207C SUPERSCRIPT EQUALS SIGN
208A SUBSCRIPT PLUS SIGN
208C SUBSCRIPT EQUALS SIGN
2100 **ACCOUNT OF**
2101 **ADDRESSED TO THE SUBJECT**
2105 **CARE OF**
2106 **CADA UNA**

3.5.3 Characters that look like characters from URLs

The following are prohibited because they look indistinguishable from the characters listed in section 3.5.1.

037E GREEK QUESTION MARK
0589 **ARMENIAN FULL STOP**
060C ARABIC COMMA
061B ARABIC SEMICOLON
066A ARABIC PERCENT SIGN
201A SINGLE LOW-9 QUOTATION MARK
2030 **PER MILLE SIGN**
2031 **PER TEN THOUSAND SIGN**
2033 **DOUBLE PRIME**
2039 **SINGLE LEFT-POINTING ANGLE QUOTATION MARK**
2044 **FRACTION SLASH**
203A SINGLE RIGHT-POINTING ANGLE QUOTATION MARK
203D INTERROBANG

3001 **IDEOGRAPHIC COMMA**
3002 **IDEOGRAPHIC FULL STOP**
3003 **DITTO MARK**
3008 **LEFT ANGLE BRACKET**
3009 **RIGHT ANGLE BRACKET**
3014 **LEFT TORTOISE SHELL BRACKET**
3015 **RIGHT TORTOISE SHELL BRACKET**
301A LEFT WHITE SQUARE BRACKET
301B RIGHT WHITE SQUARE BRACKET

3.5.4 Other punctuation

The following punctuation are prohibited because they are unlikely to be used in names and may be confusing to users or to character-entry processes:

005C REVERSE SOLIDUS

3.6 prohib-6: Symbols

[UniData] has non-normative categories for symbols. The four symbol categories are:

Symbol, Currency: Currency symbols could appear in company names and spoken phrases, so they are not prohibited.

Symbol, Modifier: Stand-alone modifiers might appear in personal names, company names, and spoken phrases, so they are not prohibited.

Symbol, Math: It is very unlikely that there are any significant personal names, company names, or spoken phrases that contain mathematical symbols. Further, many of these symbols are the same or similar to other punctuation, thereby leading to ambiguity. For this reason, math-specific symbols are prohibited. These prohibited math symbols are:

00AC NOT SIGN
00B1 PLUS-MINUS SIGN
2200-22FF [MATHEMATICAL OPERATORS]

Further, the following characters canonicalize to characters in the above math list, and therefore are also prohibited:

00BC VULGAR FRACTION ONE QUARTER
00BD VULGAR FRACTION ONE HALF
00BE VULGAR FRACTION THREE QUARTERS
207B SUPERSCRIPT MINUS
208B SUBSCRIPT MINUS
2153 **VULGAR FRACTION ONE THIRD**
2154 **VULGAR FRACTION TWO THIRDS**
2155 **VULGAR FRACTION ONE FIFTH**
2156 **VULGAR FRACTION TWO FIFTHS**

2157 **VULGAR FRACTION THREE FIFTHS**
2158 **VULGAR FRACTION FOUR FIFTHS**
2159 **VULGAR FRACTION ONE SIXTH**
215A VULGAR FRACTION FIVE SIXTHS
215B VULGAR FRACTION ONE EIGHTH
215C VULGAR FRACTION THREE EIGHTHS
215D VULGAR FRACTION FIVE EIGHTHS
215E VULGAR FRACTION SEVEN EIGHTHS
215F FRACTION NUMERATOR ONE
33A7 SQUARE M OVER S
33A8 SQUARE M OVER S SQUARED
33AE SQUARE RAD OVER S
33AF SQUARE RAD OVER S SQUARED
33C6 SQUARE C OVER KG

Symbol, Other: This category covers a multitude of symbols, few of which would ever appear in personal names, company names, and spoken phrases. The rest of the prohibited symbols are:

2190-21FF [ARROWS]
2300-23FF [MISCELLANEOUS TECHNICAL]
2400-243F [CONTROL PICTURES]
2440-245F [OPTICAL CHARACTER RECOGNITION]
2500-257F [BOX DRAWING]
2580-259F [BLOCK ELEMENTS]
25A0-25FF [GEOMETRIC SHAPES]
2600-267F [MISCELLANEOUS SYMBOLS]
2700-27BF [DINGBATS]
2800-287F [BRAILLE PATTERNS]

3.7 Additional prohibited characters

3.7.1 Unassigned characters

All characters not yet assigned in [ISO10646] are prohibited. Although this may at first seem trivial, it is extremely important because characters that may be assigned in the future might have properties that would cause them to be prohibited or might have case-folding properties. As is the case of all prohibited characters, if a DNS server receives a request containing an unassigned character, then the IDN protocol MUST return an error message.

3.7.2 Surrogate characters

So far, all proposals for binary encodings of internationalized name parts have specified UTF-8 as the encoding format. In such an encoding, surrogate characters MUST NOT be used. Therefore, for UTF-8 encodings, the following are prohibited:

D800-DFFF [SURROGATE CHARACTERS]

3.7.3 Uppercase characters with no lowercase mappings

There are many uppercase characters in [ISO10646] which do not have lowercase equivalents in [UniData]. Therefore, they are prohibited on input because they would get through the case mapping step while still being in uppercase.

The characters that are prohibited on input because they are uppercase but have no lowercase mappings are:

```
03D2        GREEK UPSILON WITH HOOK SYMBOL
03D3        GREEK UPSILON WITH ACUTE AND HOOK SYMBOL
03D4        GREEK UPSILON WITH DIAERESIS AND HOOK SYMBOL
04C0        CYRILLIC LETTER PALOCHKA
10A0-10C5   [GEORGIAN CAPITAL LETTERS]
```

Note that many characters in the range U+1200 to U+213A, the letterlike symbols, also are uppercase but have no lowercase mappings. However, they are not listed here because the entire range is already prohibited in section 3.6.

3.7.4 Radicals and Ideographic Description

Some Han characters can be informally defined in terms of ideographic descriptions. However, ideographic descriptions can lead to multiple character streams leading to the same character in a fashion that does not canonicalize. Thus, the radicals for ideographic description and the ideographic description characters themselves are prohibited. These characters are:

```
2E80-2EFF   [CJK RADICALS SUPPLEMENT]
2F00-2FDF   [KANGXI RADICALS]
2FF0-2FFF   [IDEOGRAPHIC DESCRIPTION CHARACTERS]
```

3.8 Summary of prohibited characters

The following is a collected list from the previous sections.

```
0000-001F   [CONTROL CHARACTERS]
0020        SPACE
0022        QUOTATION MARK
0023        NUMBER SIGN
0024        DOLLAR SIGN
0025        PERCENT SIGN
0026        AMPERSAND
002B        PLUS SIGN
002C        COMMA
002E        FULL STOP
002E        FULL STOP
002F        SOLIDUS
003A        COLON
003B        SEMICOLON
003C        LESS-THAN SIGN
```

003D EQUALS SIGN
003E GREATER-THAN SIGN
003F QUESTION MARK
0040 COMMERCIAL AT
005B LEFT SQUARE BRACKET
005C REVERSE SOLIDUS
005D RIGHT SQUARE BRACKET
007F DELETE
0080-009F [CONTROL CHARACTERS]
00A0 NO-BREAK SPACE
00AC NOT SIGN
00AD SOFT HYPHEN
00B1 PLUS-MINUS SIGN
00BC VULGAR FRACTION ONE QUARTER
00BD VULGAR FRACTION ONE HALF
00BE VULGAR FRACTION THREE QUARTERS
00D7 MULTIPLICATION SIGN
01C3 LATIN LETTER RETROFLEX CLICK
02B0-02FF [SPACING MODIFIER LETTERS]
037E GREEK QUESTION MARK
037E GREEK QUESTION MARK
03D2 GREEK UPSILON WITH HOOK SYMBOL
03D3 GREEK UPSILON WITH ACUTE AND HOOK SYMBOL
03D4 GREEK UPSILON WITH DIAERESIS AND HOOK SYMBOL
04C0 CYRILLIC LETTER PALOCHKA
0589 ARMENIAN FULL STOP
060C ARABIC COMMA
061B ARABIC SEMICOLON
066A ARABIC PERCENT SIGN
066D ARABIC FIVE POINTED STAR
06D4 ARABIC FULL STOP
070F SYRIAC ABBREVIATION MARK
10A0-10C5 [GEORGIAN CAPITAL LETTERS]
1680 OGHAM SPACE MARK
1806 MONGOLIAN TODO SOFT HYPHEN
180B MONGOLIAN FREE VARIATION SELECTOR ONE
180C MONGOLIAN FREE VARIATION SELECTOR TWO
180D MONGOLIAN FREE VARIATION SELECTOR THREE
180E MONGOLIAN VOWEL SEPARATOR
2000-200B [SPACES]
200C ZERO WIDTH NON-JOINER
200D ZERO WIDTH JOINER
200E LEFT-TO-RIGHT MARK
200F RIGHT-TO-LEFT MARK
2010 HYPHEN
2011 NON-BREAKING HYPHEN
2012 FIGURE DASH
2013 EN DASH
2014 EM DASH
201A SINGLE LOW-9 QUOTATION MARK
2024 ONE DOT LEADER

2025 **TWO DOT LEADER**
2026 **HORIZONTAL ELLIPSIS**
2028 **LINE SEPARATOR**
2029 **PARAGRAPH SEPARATOR**
202A LEFT-TO-RIGHT EMBEDDING
202B RIGHT-TO-LEFT EMBEDDING
202C POP DIRECTIONAL FORMATTING
202D LEFT-TO-RIGHT OVERRIDE
202E RIGHT-TO-LEFT OVERRIDE
202F NARROW NO-BREAK SPACE
2030 **PER MILLE SIGN**
2031 **PER TEN THOUSAND SIGN**
2033 **DOUBLE PRIME**
2039 **SINGLE LEFT-POINTING ANGLE QUOTATION MARK**
203A SINGLE RIGHT-POINTING ANGLE QUOTATION MARK
203D INTERROBANG
2044 **FRACTION SLASH**
2048 **QUESTION EXCLAMATION MARK**
2049 **EXCLAMATION QUESTION MARK**
206A INHIBIT SYMMETRIC SWAPPING
206B ACTIVATE SYMMETRIC SWAPPING
206C INHIBIT ARABIC FORM SHAPING
206D ACTIVATE ARABIC FORM SHAPING
206E NATIONAL DIGIT SHAPES
206F NOMINAL DIGIT SHAPES
207A SUPERSCRIPT PLUS SIGN
207B SUPERSCRIPT MINUS
207C SUPERSCRIPT EQUALS SIGN
208A SUBSCRIPT PLUS SIGN
208B SUBSCRIPT MINUS
208C SUBSCRIPT EQUALS SIGN
2100 **ACCOUNT OF**
2101 **ADDRESSED TO THE SUBJECT**
2105 **CARE OF**
2106 **CADA UNA**
2153 **VULGAR FRACTION ONE THIRD**
2154 **VULGAR FRACTION TWO THIRDS**
2155 **VULGAR FRACTION ONE FIFTH**
2156 **VULGAR FRACTION TWO FIFTHS**
2157 **VULGAR FRACTION THREE FIFTHS**
2158 **VULGAR FRACTION FOUR FIFTHS**
2159 **VULGAR FRACTION ONE SIXTH**
215A VULGAR FRACTION FIVE SIXTHS
215B VULGAR FRACTION ONE EIGHTH
215C VULGAR FRACTION THREE EIGHTHS
215D VULGAR FRACTION FIVE EIGHTHS
215E VULGAR FRACTION SEVEN EIGHTHS
215F FRACTION NUMERATOR ONE
2160-217F [ROMAN NUMERALS]
2190-21FF [ARROWS]
2200-22FF [MATHEMATICAL OPERATORS]

2300-23FF [MISCELLANEOUS TECHNICAL]
2400-243F [CONTROL PICTURES]
2440-245F [OPTICAL CHARACTER RECOGNITION]
2488 DIGIT ONE FULL STOP
2489 DIGIT TWO FULL STOP
248A DIGIT THREE FULL STOP
248B DIGIT FOUR FULL STOP
248C DIGIT FIVE FULL STOP
248D DIGIT SIX FULL STOP
248E DIGIT SEVEN FULL STOP
248F DIGIT EIGHT FULL STOP
2490 DIGIT NINE FULL STOP
2491 NUMBER TEN FULL STOP
2492 NUMBER ELEVEN FULL STOP
2493 NUMBER TWELVE FULL STOP
2494 NUMBER THIRTEEN FULL STOP
2495 NUMBER FOURTEEN FULL STOP
2496 NUMBER FIFTEEN FULL STOP
2497 NUMBER SIXTEEN FULL STOP
2498 NUMBER SEVENTEEN FULL STOP
2499 NUMBER EIGHTEEN FULL STOP
249A NUMBER NINETEEN FULL STOP
249B NUMBER TWENTY FULL STOP
2500-257F [BOX DRAWING]
2580-259F [BLOCK ELEMENTS]
25A0-25FF [GEOMETRIC SHAPES]
2600-267F [MISCELLANEOUS SYMBOLS]
2700-27BF [DINGBATS]
2800-287F [BRAILLE PATTERNS]
2E80-2EFF [CJK RADICALS SUPPLEMENT]
2F00-2FDF [KANGXI RADICALS]
2FF0-2FFF [IDEOGRAPHIC DESCRIPTION CHARACTERS]
3000 IDEOGRAPHIC SPACE
3001 IDEOGRAPHIC COMMA
3002 IDEOGRAPHIC FULL STOP
3003 DITTO MARK
3008 LEFT ANGLE BRACKET
3009 RIGHT ANGLE BRACKET
33A7 SQUARE M OVER S
33A8 SQUARE M OVER S SQUARED
33AE SQUARE RAD OVER S
33AF SQUARE RAD OVER S SQUARED
33C2 SQUARE AM
33C2 SQUARE AM
33C6 SQUARE C OVER KG
33C7 SQUARE CO
33D8 SQUARE PM
33D8 SQUARE PM
D800-DFFF [SURROGATE CHARACTERS]
E000-F8FF [PRIVATE USE, PLANE 0]
FB1D-FB4F [HEBREW PRESENTATION FORMS]

FB50-FDFF [ARABIC PRESENTATION FORMS A]
FE20-FE2F [COMBINING HALF MARKS]
FE30-FE4F [CJK COMPATIBILITY FORMS]
FE50-FE6F [SMALL FORM VARIANTS]
FE70-FEFC [ARABIC PRESENTATION FORMS B]
FEFF ZERO WIDTH NO-BREAK SPACE
FF00-FFEF [HALFWIDTH AND FULLWIDTH FORMS]
FFF9 INTERLINEAR ANNOTATION ANCHOR
FFFA INTERLINEAR ANNOTATION SEPARATOR
FFFB INTERLINEAR ANNOTATION TERMINATOR
FFFC OBJECT REPLACEMENT CHARACTER
FFFD REPLACEMENT CHARACTER
Unassigned characters

4. Case Folding

After it has been verified that the input text has none of the characters prohibited for case folding, the case-folding step itself is quite straight-forward. For each character in the input, if there is a lowercase mapping for that character in [UniData], the input character is changed to the mapped lowercase letter.

5. Canonicalization

After case folding, the input string is normalized using form KC, as described in [UTR15].

6. IDN Table Revisions

A table consisting of all characters allowed and prohibited and the rules for case folding and canonicalization will be created based on the content of the [UniData] and on the content of this document. This table will be the authority for implementations to follow and will be normatively referenced by this document. Such a table will enable the IDN protocol to have versions independent of the revisions to Unicode and/or to ISO 10646 because the revision of IDN and its deployment may not in sync with revisions to Unicode and ISO 10646.

In a future draft of this document, IANA will be asked to keep this table, with an initial version number of 1. Each new version of the table will have a new, higher version number.

7. 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.

Host names 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 host name.

8. References

[IDNComp] Paul Hoffman, "Comparison of Internationalized Domain Name Proposals", draft-ietf-idn-compare.

[IDNReq] James Seng, "Requirements of Internationalized Domain Names", draft-ietf-idn-requirement.

[ISO10646] ISO/IEC 10646-1:1993. International Standard -- Information technology -- Universal Multiple-Octet Coded Character Set (UCS) -- Part 1: Architecture and Basic Multilingual Plane. Five amendments and a technical corrigendum have been published up to now. UTF-16 is described in Annex Q, published as Amendment 1. 17 other amendments are currently at various stages of standardization. [[[THIS REFERENCE NEEDS TO BE UPDATED AFTER DETERMINING ACCEPTABLE WORDING]]]

[Normalize] Character Normalization in IETF Protocols, draft-duerst-i18n-norm-03

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

[RFC2396] Tim Berners-Lee, et. al., "Uniform Resource Identifiers (URI): Generic Syntax", August 1998, RFC 2396.

[RFC2732] Robert Hinden, et. al., Format for Literal IPv6 Addresses in URL's, December 1999, RFC 2732.

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

[Unicode3] The Unicode Consortium, "The Unicode Standard -- Version 3.0", ISBN 0-201-61633-5. Described at <http://www.unicode.org/unicode/standard/versions/Unicode3.0.html>.

[UniData] The Unicode Consortium. UnicodeData File. <ftp://ftp.unicode.org/Public/UNIDATA/UnicodeData.txt>.

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

A. Acknowledgements

Many people from the IETF IDN Working Group and the Unicode Technical Committee contributed ideas that went into the first draft of this

document. Mark Davis was particularly helpful in some of the early ideas.

B. Changes From Previous Versions of this Draft

This is the -00 version, so there are no changes.

C. IANA Considerations

There are no specific IANA considerations in this draft, but there will be in a future draft of this document.

D. Author Contact Information

Paul Hoffman
Internet Mail Consortium and VPN Consortium
127 Segre Place
Santa Cruz, CA 95060 USA
paul.hoffman@imc.org and paul.hoffman@vpnc.org

Marc Blanchet
Viagenie inc.
2875 boul. Laurier, bur. 300
Ste-Foy, Quebec, Canada, G1V 2M2
Marc.Blanchet@viagenie.qc.ca