L2/02-020

Internet Draft
draft-ietf-idn-nameprep-07.txt
January 9, 2001
Expires in six months

Paul Hoffman
IMC & VPNC
Marc Blanchet
ViaGenie

Stringprep Profile for Internationalized Host Names

Status of this memo

Abstract

This document describes how to prepare internationalized host name parts in order to increase the likelihood that name input and name comparison work in ways that make sense for typical users throughout the world. This profile of the stringprep protocol is used as part of a suite of on-the-wire protocols for internationalizing the DNS.

1. Introduction

This document specifies processing rules that will allow users to enter internationalized host name parts in applications and have the highest chance of getting the content of the strings correct. It is a profile of stringprep [STRINGPREP].

This document was previously called "nameprep" before splitting the structure of the protocol off into the stringprep document.

This profile defines the following, as required by [STRINGPREP]

- The intended applicability of the profile: internationalized host name parts

- The character repertoire that is the input and output to stringprep: defined in Section 2

- The list of unassigned code points for the repertoire: defined in Appendix F.

- The mappings used: defined in Section 3.

- The Unicode normalization used: defined in Section 4

- The characters that are prohibited as output: Defined in section 5

1.2 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 for code points and names from the Unicode Standard [Unicode3.1] and ISO/IEC 10646 [ISO10646]. 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. The comments for character ranges are shown in square brackets (such as "[SYMBOLS]") and do not come from the standards.

2. Character Repertoire

Unicode 3.1 [Unicode3.1] is the repertoire used in this profile. The reason Unicode 3.1 was chosen instead of a version of ISO/IEC 10646 is that ISO/IEC 10646 is expected to be updated soon after this document becomes an RFC. Unicode 3.1 has the exact repertoire that is expected in the next version of ISO/IEC 10646, and is therefore used here.

3. Mapping

This profile specifies stringprep mapping using the mapping table in Appendix D. That table includes all the steps described in this section.

Note that text in this section describe how Appendix D was formed. It is there for people who want to understand more, but it should be ignored by implementors. Implementations of this profile MUST map based on Appendix D, not based on the descriptions in this section of how Appendix D was created.

3.1 Mapped out

The following characters are simply deleted from the input (that is, they are mapped to nothing) because their presence or absence should not make two strings different.

Some characters are only useful in line-based text, and are otherwise invisible and ignored.

00AD; SOFT HYPHEN
1806; MONGOLIAN TODO SOFT HYPHEN
200B; ZERO WIDTH SPACE
FEFF; ZERO WIDTH NO-BREAK SPACE

Variation selectors and cursive connectors select different glyphs, but do not bear semantics.

180B; MONGOLIAN FREE VARIATION SELECTOR ONE
180C; MONGOLIAN FREE VARIATION SELECTOR TWO
180D; MONGOLIAN FREE VARIATION SELECTOR THREE
200C; ZERO WIDTH NON-JOINER
200D; ZERO WIDTH JOINER

3.2 Case mapping

The input string is case folded according to [UTR21]. For most characters, this is the same as changing the input character to a lowercase character. For some characters, however, more complex transformations occur. The "CaseFolding.txt" file from the Unicode database was used to prepare the mapping table.

There are some characters that do not have mappings in [UTR21] but still need processing. These characters include a few Greek characters and many symbols that contain Latin characters. The list of characters to add to the mapping table were determined by the following algorithm:

```
b = NormalizeWithKC(Fold(a));
c = NormalizeWithKC(Fold(b));
if c is not the same as b, add a mapping for "a to c".
```

Because NormalizeWithKC(Fold(c)) always equals c, the table is stable from that point on. The "DerivedNormalizationProperties.txt" file from the Unicode database was used to prepare Appendix D. This mapping was added to reduce the number of processing steps, that is, to avoid doing case mapping and normalization twice.

4. Normalization

This profile specifies using Unicode normalization form KC, as described in [UAX15].

5. Prohibited Output

This profile specifies using the prohibition table in Appendix E.

Note that the subsections below describe how Appendix E was formed. They are there for people who want to understand more, but they should be ignored by implementors. Implementations of this profile MUST map based on Appendix E, not based on the descriptions in this section of how Appendix E was created.

The collected lists of prohibited code points can be found in Appendix E of this document. The lists in Appendix E MUST be used by implementations of this specification. If there are any discrepancies between the lists in Appendix E and subsections below, the lists in Appendix E always takes precedence.

Some code points listed in one section would also appear in other sections. Each code point is only listed once in the tables in Appendix E.

5.1 Space characters

Space characters would make visual transcription of URLs nearly impossible and could lead to user entry errors in many ways.

0020; SPACE
00A0; NO-BREAK SPACE
1680; OGHAM SPACE MARK
2000; EN QUAD
2001; EM QUAD
2002; EN SPACE
2003; EM SPACE
2004; THREE-PER-EM SPACE
2005; FOUR-PER-EM SPACE
2006; SIX-PER-EM SPACE
2007; FIGURE SPACE
2008; PUNCTUATION SPACE
2009; THIN SPACE
200A; HAIR SPACE
202F; NARROW NO-BREAK SPACE
3000; IDEOGRAPHIC SPACE

5.2 Control characters

Control characters (or characters with control function) cannot be seen and can cause unpredictable results when displayed.

0000-001F; [CONTROL CHARACTERS]
007F; DELETE
0080-009F; [CONTROL CHARACTERS]
070F; SYRIAC ABBREVIATION MARK
180E; MONGOLIAN VOWEL SEPARATOR
2028; LINE SEPARATOR
2029; PARAGRAPH SEPARATOR
206A-206F; [CONTROL CHARACTERS]
FFF9-FFFC; [CONTROL CHARACTERS]
1D173-1D17A; [MUSICAL CONTROL CHARACTERS]

5.3 Private use and replacement characters

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

E000-F8FF; [PRIVATE USE, PLANE 0]
F0000-FFFFD; [PRIVATE USE, PLANE 15]
100000-10FFFD; [PRIVATE USE, PLANE 16]

The replacement character (U+FFFD) has no known semantic definition in a name, and is often displayed by renderers to indicate "there would be some character here, but it cannot be rendered". For example, on a computer with no Asian fonts, a name with three ideographs might be rendered with three replacement characters.

FFFD; REPLACEMENT CHARACTER

5.4 Non-character code points

Non-character code points are code points that have been allocated in ISO/IEC 10646 but are not characters. Because they are already assigned, they are guaranteed not to later change into characters.

FDD0-FDEF; [NONCHARACTER CODE POINTS]
FFFE-FFFF; [NONCHARACTER CODE POINTS]
1FFFE-1FFFF; [NONCHARACTER CODE POINTS]
2FFFE-2FFFF; [NONCHARACTER CODE POINTS]
3FFFE-3FFFF; [NONCHARACTER CODE POINTS]
4FFFE-4FFFF; [NONCHARACTER CODE POINTS]
5FFFE-5FFFF; [NONCHARACTER CODE POINTS]
6FFFE-6FFFF; [NONCHARACTER CODE POINTS]
7FFFE-7FFFF; [NONCHARACTER CODE POINTS]
8FFFE-8FFFF; [NONCHARACTER CODE POINTS]
9FFFE-9FFFF; [NONCHARACTER CODE POINTS]
AFFFE-AFFFF; [NONCHARACTER CODE POINTS]
BFFFE-BFFFF; [NONCHARACTER CODE POINTS]
CFFFE-CFFFF; [NONCHARACTER CODE POINTS]
DFFFE-DFFFF; [NONCHARACTER CODE POINTS]
EFFFE-EFFFF; [NONCHARACTER CODE POINTS]
FFFFE-FFFFF; [NONCHARACTER CODE POINTS]
10FFFE-10FFFF; [NONCHARACTER CODE POINTS]

The non-character code points are listed the PropList.txt file from the Unicode database.

5.5 Surrogate codes

The following code points are permanently reserved for use as surrogate code values in the UTF-16 encoding, will never be assigned to characters, and are therefore prohibited:

D800-DFFF; [SURROGATE CODES]

5.6 Inappropriate for plain text

The following characters should not appear in regular text.

FFF9; INTERLINEAR ANNOTATION ANCHOR
FFFA; INTERLINEAR ANNOTATION SEPARATOR
FFFB; INTERLINEAR ANNOTATION TERMINATOR
FFFC; OBJECT REPLACEMENT CHARACTER

5.7 Inappropriate for canonical representation

The ideographic description characters allow different sequences of characters to be rendered the same way, which makes them inappropriate for host names that must have a single canonical representation.

2FF0-2FFB; [IDEOGRAPHIC DESCRIPTION CHARACTERS]

5.8 Change display properties

The following characters, some of which are deprecated in ISO/IEC 10646, can cause changes in display or the order in which characters appear when rendered.

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

5.9 Inappropriate characters from common input mechanisms

U+3002 is used as if it were U+002E in many input mechanisms, particularly in Asia. This prohibition allows input mechanisms to safely map U+3002 to U+002E before doing stringprep without worrying about preventing users from accessing legitimate host name parts.

3002; IDEOGRAPHIC FULL STOP

5.10 Tagging characters

The following characters are used for tagging text and are invisible.

E0001; LANGUAGE TAG
E0020-E007F; [TAGGING CHARACTERS]

6. Unassigned Code Points in Internationalized Host Names

This profile lists the unassigned code points for Unicode 3.1 in Appendix F. The list in Appendix F MUST be used by implementations of this specification. If there are any discrepancies between the list in Appendix F and the Unicode 3.1 specification, the list Appendix F always takes precedence.

7. Security Considerations

ISO/IEC 10646 has many characters that look similar. In many cases, users of security protocols might do visual matching, such as when comparing the names of trusted third parties. This profile does nothing to map similar-looking characters together.

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.

Current applications may assume that the characters allowed in host

names will always be the same as they are in [STD13]. This document vastly increases the number of characters available in host names. Every program that uses "special" characters in conjunction with host names may be vulnerable to attack based on the new characters allowed by this specification.

8. References

[CharModel] Unicode Technical Report;17, Character Encoding Model. <http://www.unicode.org/unicode/reports/tr17/>.

[Glossary] Unicode Glossary, <http://www.unicode.org/glossary/>.

[ISO10646] ISO/IEC 10646-1:2000. International Standard -- Information technology -- Universal Multiple-Octet Coded Character Set (UCS) - Part 1: Architecture and Basic Multilingual Plane.

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

[STD13] Paul Mockapetris, "Domain names - concepts and facilities" (RFC 1034) and "Domain names - implementation and specification" (RFC 1035, STD 13, November 1987.

[STRINGPREP] Paul Hoffman and Marc Blanchet, "Preparation of Internationalized Strings ("stringprep")", draft-hoffman-stringprep, work in progress

[Unicode3.1] The Unicode Standard, Version 3.1.0: The Unicode Consortium. The Unicode Standard, Version 3.0. Reading, MA, Addison-Wesley Developers Press, 2000. ISBN 0-201-61633-5, as amended by: Unicode Standard Annex #27: Unicode 3.1 <http://www.unicode.org/unicode/reports/tr27/tr27-4.html>.

[URI] For example: Roy Fielding et al., "Uniform Resource Identifiers: Generic Syntax", August 1998, RFC 2396; Robert Hinden et. al, "IPv6 Literal Addresses in URL's", December 1999, RFC 2732. Note that there are many other RFCs that define additional URI schemes.

[UAX15] Mark Davis and Martin Duerst. Unicode Standard Annex #15: Unicode Normalization Forms, Version 3.1.0. <http://www.unicode.org/unicode/reports/tr15/tr15-21.html>

[UTR21] Mark Davis. Case Mappings. Unicode Technical Report;21. <http://www.unicode.org/unicode/reports/tr21/>.

9. Differences Between -06 and -07 Drafts

5: Removed 5.1 (currently-used ASCII characters) and renumbered the entire section.

E: Removed the characters that appeared in the old 5.1.

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.

The IDN namprep design team made many useful changes to the first draft. That team and its advisors include:

Asmus Freytag
Cathy Wissink
Francois Yergeau
James Seng
Marc Blanchet
Mark Davis
Martin Duerst
Patrik Faltstrom
Paul Hoffman

Additional significant improvements were proposed by:

Jonathan Rosenne
Kent Karlsson
Scott Hollenbeck
Dave Crocker

B. IANA Considerations

This is a profile of stringprep. When it becomes an RFC, it should be registered in the stringprep profile registry.

C. 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

THE TABLES HAVE BEEN REMOVED FOR L2 DISTRIBUTION, AND MAY BE FOUND ON THE WEB AT:
http://www.ietf.org/internet-drafts/draft-gaisher-geostd8-00.txt
http://www.unicode.org/L2/L2002/02020-tables.txt