Abstract

This document describes how to prepare UTF-8 strings for use with Kerberos protocols 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 is a profile of "Preparation of Internationalized Strings" [RFC3454].

1. Introduction

This document specifies processing rules that will allow users to enter Kerberos Principal Names and input to cryptographic String to Key functions. It is a profile of stringprep [RFC3454].

This profile defines the following, as required by [RFC3454]

- 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 D.
- 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 3.2 [Unicode3.2] is the repertoire used in this profile. The reason Unicode 3.2 was chosen instead of a version of ISO/IEC 10646 is that Unicode 3.2 is the basis for [RFC3454].

3. Mapping

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

Note that text in this section describe how Appendix C 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 C, not based on the descriptions in this section of how Appendix C was created.

3.1 Mapped to nothing

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

3.2 Space Character Conversions

Space characters can make accurate visual transcription of names nearly impossible and could lead to user entry errors in many ways. The following Unicode spaces are to be mapped to 0020; SPACE:

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
4. Normalization

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

NOTE: There was some discussion on the mailing list that would suggest that Unicode NFKC does not properly handle the composition of normalized Hangul strings. Following the lead of the IDN working group, the Kerberos working group will not attempt to second-guess the authors of Unicode 3.1 Annex 15 (formerly Technical Report 15) [UAX15], which specifies the normalization methods, or the Ideographic Rappoporteur Group (IRG), which is the formal subgroup of ISO/IEC JTC1/SC2/WG2 charged with approving all CJKV elements of the Unicode standards. Such issues are outside the working group's charter and its area of expertise.

5. Prohibited Output

This profile specifies using the prohibition table in Appendix D.

Note that the subsections below describe how Appendix D 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 D, not based on the descriptions in this section of how Appendix D was created.

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

5.1 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]
06DD; ARABIC END OF AYAH
5.2 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]

5.3 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]
1FFFFE–1FFFFF; [NONCHARACTER CODE POINTS]
2FFFE–2FFFFF; [NONCHARACTER CODE POINTS]
3FFFE–3FFFFF; [NONCHARACTER CODE POINTS]
4FFFE–4FFFFF; [NONCHARACTER CODE POINTS]
5FFFE–5FFFFF; [NONCHARACTER CODE POINTS]
6FFFE–6FFFFF; [NONCHARACTER CODE POINTS]
7FFFE–7FFFFF; [NONCHARACTER CODE POINTS]
8FFFE–8FFFFF; [NONCHARACTER CODE POINTS]
9FFFE–9FFFFF; [NONCHARACTER CODE POINTS]
AFFFE–AFFFFE; [NONCHARACTER CODE POINTS]
BFFFE–BFFFFF; [NONCHARACTER CODE POINTS]
CFFFE–CFFFFF; [NONCHARACTER CODE POINTS]
DFFFE–DFFFFF; [NONCHARACTER CODE POINTS]
EFFFE–EFFFFE; [NONCHARACTER CODE POINTS]
FFFFE–FFFFFF; [NONCHARACTER CODE POINTS]
10FFFE–10FFFFF; [NONCHARACTER CODE POINTS]

The non-character code points are listed the PropList.txt file from the Unicode database.
5.4 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]

5.5 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

Although the replacement character (U+FFFD) might be used when a name is
displayed, it doesn't make sense for it to be part of the name itself.
It 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

5.6 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]

5.7 Change display properties

The following characters can cause changes in display or the order in
which characters appear when rendered, or are deprecated in Unicode.

0340; COMBINING GRAVE TONE MARK
0341; COMBINING ACUTE TONE MARK
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
5.8 Tagging characters

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

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

6. Bidirectional Characters

This profile specifies checking bidirectional strings as described in [RFC3454] section 6.

7. Unassigned Code Points

This profile lists the unassigned code points for Unicode 3.2 in Appendix E. The list in Appendix E MUST be used by implementations of this specification. If there are any discrepancies between the list in Appendix E and the Unicode 3.2 specification, the list Appendix E always takes precedence.

8. 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.

Principal names and passwords are entered by users and used within the Kerberos protocol. The security of the Internet would be compromised if a user entering a single internationalized string could be connected to different servers or denied access based on different interpretations of internationalized strings.

9. IANA Considerations

IANA is to register this profile as described in [RFC3454].

10. References


A. Acknowledgements

This draft is based upon the work of the IETF IDN Working Group's
IDN Nameprep design team.

This profile is the work of the Kerberos Working Group. Significant
contributions were provided by Jeffrey Hutzelman, Sam Hartman, Tom Yu,
Ken Raeburn, and Jeffrey Altman.

B. Editor Contact Information

Jeffrey Altman
Internet Access Methods
200 W 72ND ST STE 48
NEW YORK NY 10024
+1 (212) 580-2700
e-mail: jaltman@iamx.com

C. Mapping Tables

The following is the mapping table from Section 3. The table has three
columns:
- the character that is mapped from
- the zero or more characters that it is mapped to
- the reason for the mapping

The columns are separated by semicolons. Note that the second column may be empty, or it may have one character, or it may have more than one character, with each character separated by a space.

<table>
<thead>
<tr>
<th>Character</th>
<th>Mapping Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>00A0</td>
<td>NO-BREAK SPACE</td>
</tr>
<tr>
<td>00AD</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>034F</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>1680</td>
<td>OGHAM SPACE MARK</td>
</tr>
<tr>
<td>1806</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>180B</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>180C</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>180D</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>2000</td>
<td>EN QUAD</td>
</tr>
<tr>
<td>2001</td>
<td>EM QUAD</td>
</tr>
<tr>
<td>2002</td>
<td>EN SPACE</td>
</tr>
<tr>
<td>2003</td>
<td>EM SPACE</td>
</tr>
<tr>
<td>2004</td>
<td>THREE-PER-EM SPACE</td>
</tr>
<tr>
<td>2005</td>
<td>FOUR-PER-EM SPACE</td>
</tr>
<tr>
<td>2006</td>
<td>SIX-PER-EM SPACE</td>
</tr>
<tr>
<td>2007</td>
<td>FIGURE SPACE</td>
</tr>
<tr>
<td>2008</td>
<td>PUNCTUATION SPACE</td>
</tr>
<tr>
<td>2009</td>
<td>THIN SPACE</td>
</tr>
<tr>
<td>200A</td>
<td>HAIR SPACE</td>
</tr>
<tr>
<td>200B</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>200C</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>200D</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>2060</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>202F</td>
<td>NARROW NO-BREAK SPACE</td>
</tr>
<tr>
<td>205F</td>
<td>MEDIUM MATHEMATICAL SPACE</td>
</tr>
<tr>
<td>3000</td>
<td>IDEOGRAPHIC SPACE</td>
</tr>
<tr>
<td>FE00</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FE01</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FE02</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FE03</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FE04</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FE05</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FE06</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FE07</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FE08</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FE09</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FE0A</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FE0B</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FE0C</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FE0D</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FE0E</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FE0F</td>
<td>Map to nothing</td>
</tr>
<tr>
<td>FEFF</td>
<td>Map to nothing</td>
</tr>
</tbody>
</table>
D. Prohibited Code Point List

----- Start Prohibited Table -----  
0000-001F; [CONTROL CHARACTERS]  
007F; DELETE  
0080-009F; [CONTROL CHARACTERS]  
0340; COMBINING GRAVE TONE MARK  
0341; COMBINING ACUTE TONE MARK  
06DD; ARABIC END OF AYAH  
070F; SYRIAC ABBREVIATION MARK  
100000-10FFFFD; [PRIVATE USE, PLANE 16]  
10FFFFE-10FFFFF; [NONCHARACTER CODE POINTS]  
180E; MONGOLIAN VOWEL SEPARATOR  
1D173-1D17A; [MUSICAL CONTROL CHARACTERS]  
1FFFE-1FFFF; [NONCHARACTER CODE POINTS]  
200C; ZERO WIDTH NON-JOINER  
200D; ZERO WIDTH JOINER  
200E; LEFT-TO-RIGHT MARK  
200F; RIGHT-TO-LEFT MARK  
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  
2060; WORD JOINER  
2061; FUNCTION APPLICATION  
2062; INVISIBLE TIMES  
2063; INVISIBLE SEPARATOR  
206A-206F; [CONTROL CHARACTERS]  
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  
2FF0-2FFB; [IDEOGRAPHIC DESCRIPTION CHARACTERS]  
2FFFE-2FFFF; [NONCHARACTER CODE POINTS]  
3FFFE-3FFFFF; [NONCHARACTER CODE POINTS]  
4FFFE-4FFFFF; [NONCHARACTER CODE POINTS]  
5FFFE-5FFFFF; [NONCHARACTER CODE POINTS]  
6FFFE-6FFFFF; [NONCHARACTER CODE POINTS]  
7FFFE-7FFFFF; [NONCHARACTER CODE POINTS]  
8FFFE-8FFFFF; [NONCHARACTER CODE POINTS]  
9FFFE-9FFFFF; [NONCHARACTER CODE POINTS]  
AFFFE-AFFFF; [NONCHARACTER CODE POINTS]  
BFFFE-BFFFFF; [NONCHARACTER CODE POINTS]
NOTE WELL: Software that follows this specification that will be used to check names before they are put in authoritative name servers MUST add all unassigned code points to the list of characters that are prohibited. See Section 6 of [RFC3454] for more details.

### E. Unassigned Code Point List

<table>
<thead>
<tr>
<th>Code Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>0221</td>
</tr>
<tr>
<td>0234–024F</td>
</tr>
<tr>
<td>02AE–02AF</td>
</tr>
<tr>
<td>02EF–02FF</td>
</tr>
<tr>
<td>0350–035F</td>
</tr>
<tr>
<td>0370–0373</td>
</tr>
<tr>
<td>0376–0379</td>
</tr>
<tr>
<td>037B–037D</td>
</tr>
<tr>
<td>037F–0383</td>
</tr>
<tr>
<td>038B</td>
</tr>
<tr>
<td>038D</td>
</tr>
<tr>
<td>03A2</td>
</tr>
<tr>
<td>03CF</td>
</tr>
<tr>
<td>03F7–03FF</td>
</tr>
<tr>
<td>0487</td>
</tr>
<tr>
<td>04CF</td>
</tr>
<tr>
<td>04F6–04F7</td>
</tr>
<tr>
<td>04FA–04FF</td>
</tr>
<tr>
<td>0510–0530</td>
</tr>
<tr>
<td>0557–0558</td>
</tr>
<tr>
<td>0560</td>
</tr>
<tr>
<td>0588</td>
</tr>
</tbody>
</table>
1D800-1FFFD
2A6D7-2F7FF
2FA1E-2FFFD
30000-3FFFD
40000-4FFFD
50000-5FFFD
60000-6FFFD
70000-7FFFD
80000-8FFFD
90000-9FFFD
A0000-AFFFD
B0000-BFFFD
C0000-CFFFD
D0000-DFFFD
E0000
E0002-E001F
E0080-EFFFD
----- End Unassigned Table -----