Mapping characters for PRECIS classes
draft-ietf-precis-mappings-12

Abstract

The framework for preparation, enforcement, and comparison of internationalized strings ("PRECIS") defines several classes of strings for use in application protocols. Because many protocols perform case-sensitive or case-insensitive string comparison, it necessary to define methods for case mapping. In addition, both the Internationalized Domain Names in Applications (IDNA) and the PRECIS problem statement describe mappings for internationalized strings that are not limited to case, but include width mapping and mapping of delimiters and other special characters that can be taken into consideration. This document provides guidelines for designers of PRECIS profiles and describes several mappings that can be applied between receiving user input and passing permitted code points to internationalized protocols. In particular, this document describes both locale-dependent and context-depending case mappings as well as additional mappings for delimiters and special characters.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on May 4, 2016.
1. Introduction

In many cases, user input of internationalized strings is generated through the use of an input method editor ("IME") or through copy-and-paste from free text. Users generally do not care about the case and/or width of input characters because they consider those characters to be functionally equivalent or visually identical. Furthermore, users rarely switch the IME state to input special characters such as protocol elements. For Internationalized Domain Names ("IDNs"), the IDNA Mapping specification [RFC5895] describes methods for handling these issues. For PRECIS strings, case mapping and width mapping are defined in the PRECIS framework specification [RFC7564]. The case and width mappings defined in the PRECIS framework do not handle other mappings such as delimiter characters, special characters, and locale-dependent or context-dependent case; these mappings are also important in order to increase the probability that the resulting strings compare as users expect. This document provides guidelines for authors of protocol profiles of the PRECIS framework and describes several mappings that can be applied between receiving user input and passing permitted code points to internationalized protocols. The delimiter mapping and special mapping rules described here are applied as "additional mappings" beyond those defined in the PRECIS framework, whereas the "local case mapping" rule provides locale-dependent and context-dependent alternative case mappings for specific target characters.

2. Protocol dependent mappings

The PRECIS framework defines several protocol-independent mappings. The additional mappings and local case mapping defined in this document are protocol-dependent, i.e., they depend on the rules for a particular application protocol.

2.1. Delimiter mapping

Some application protocols define delimiters for their own use, resulting in the fact that the delimiters are different for each protocol. The delimiter mapping table should therefore be based on a well-defined mapping table for each protocol.

Delimiter mapping is used to map characters that are similar to protocol delimiters into the canonical delimiter characters. For example, there are width-compatible characters that correspond to the ‘@’ in email addresses and the ‘:/’ in URIs. The ‘+’, ‘-’, ‘<’ and ‘>’ characters are other common delimiters that might require such mapping. For the FULL STOP character (U+002E), a delimiter in the visual presentation of domain names, some IMEs produce a character such as IDEOGRAPHIC FULL STOP (U+3002) when a user types
FULL STOP on the keyboard. In all these cases, the visually similar characters that can come from user input need to be mapped to the correct protocol delimiter characters before the string is passed to the protocol.

2.2. Special mapping

Aside from delimiter characters, certain protocols have characters which need to be mapped in ways that are different from the rules specified in the PRECIS framework (e.g., mapping non-ASCII space characters to ASCII space). In this document, these mappings are called "special mappings". They are different for each protocol. Therefore, the special mapping table should be based on a well-defined mapping table for each protocol. Examples of special mapping are the following:

- White spaces such as CHARACTER TABULATION(U+0009) or IDEOGRAPHIC SPACE(U+3000) are mapped to SPACE (U+0020)
- Some characters such as control characters are mapped to nothing (Deletion)

As examples, EAP [RFC3748], SASLprep [RFC4013], IMAP4 ACL [RFC4314] and LDAPprep [RFC4518] define the rule that some codepoints for the non-ASCII space are mapped to SPACE (U+0020).

2.3. Local case mapping

The purpose of local case mapping is to increase the probability of results that users expect when character case is changed (e.g., map uppercase to lowercase) between input and use in a protocol. Local case mapping selectively affects characters whose case mapping depends on locale and/or context.

(Note: The term "locale" in this document practically means "language" or "language and region" because the locale based on that language configuration of applications on POSIX is selected by "locale" information and referred "Note" in section 2.1.1 of BCP 47 [RFC5646].)

As an example of locale and context-dependent mapping, LATIN CAPITAL LETTER I ("I", U+0049) is normally mapped to LATIN SMALL LETTER I ("i", U+0069); however, if the language is Turkish (or one of several other languages), unless an I is before a dot_above, the character should be mapped to LATIN SMALL LETTER DOTLESS I (U+0131).

Case mapping using Unicode Default Case Folding in the PRECIS framework does not consider such locale or context because it is a common framework for internationalization. Local case mapping
defined in this document corresponds to demands from applications which supports users’ locale and/or context. The complete set of possible target characters for local case mapping are the characters specified in the SpecialCasing.txt [Specialcasing] file in the Section 3.13 of the Unicode Standard [Unicode], but the specific set of target characters selected for local case mapping depends on locale and/or context, as further explained in the SpecialCasing.txt file.

The case folding method for a selected target character is to map into lower case as defined in SpecialCasing.txt. The case folding method for all other, non-target characters is as specified in the Section 5.2.3 of the PRECIS framework. When an application supports users’ locale and/or context, use of local case mapping can increase the probability that string comparisons yield the results that users expect.

If a PRECIS profile selects Unicode Default Case Folding as the preferred method of case mapping, the profile designers may consider whether local case mapping can be applied. And if it can be applied, it is better to add "alternatively, local case mapping might be applicable" after "Unicode Default Case Folding" so that application developers are aware of the alternative. See the Appendix B for a description of why local case mapping can be an alternative.

3. Order of operations

Delimiter mapping and special mapping as described in this document are expected to be applied as the "Additional Mapping Rule" mentioned in the Section 5.2.2 of the PRECIS framework. Although the delimited mapping and special mapping could be applied in either order, this document recommends the following order to minimize the effect of code point changes introduced by the mappings and to be acceptable to the widest user community:

1. Delimiter mapping
2. Special mapping

4. Security Considerations

Detailed security considerations for PRECIS strings are discussed in the PRECIS framework specification [RFC7564]. This document inherits the considerations as well.

As with Mapping Characters for IDNA2008 [RFC5895], this document suggests creating mappings that might cause confusion for some users
while alleviating confusion in other users. Such confusion is not covered in any depth in this document.

5. IANA Considerations

This document has no actions for the IANA.

6. Acknowledgment

Martin Duerst suggested a need for the case folding about the mapping (map final sigma to sigma, German sz to ss, ).

Alexey Melnikov, Andrew Sullivan, Barry Leiba, David Black, Heather Flanagan, Joe Hildebrand, John Klensin, Marc Blanchet, Pete Resnick and Peter Saint-Andre, et al. gave important suggestion for this document during working group discussions.

7. References

7.1. Normative References


7.2. Informative References


Appendix A. Mapping type list each protocol

A.1. Mapping type list for each protocol

This table is the mapping type list for each protocol. Values marked "o" indicate that the protocol use the type of mapping. Values marked "-" indicate that the protocol doesn’t use the type of mapping.

<table>
<thead>
<tr>
<th>Protocol and mapping RFC</th>
<th>Width (NFKC)</th>
<th>Delimiter</th>
<th>Case</th>
<th>Special</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDNA (RFC 3490)</td>
<td>-</td>
<td>o</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IDNA (RFC 3491)</td>
<td>o</td>
<td>-</td>
<td>o</td>
<td>-</td>
</tr>
<tr>
<td>iSCSI (RFC 3722)</td>
<td>o</td>
<td>-</td>
<td>o</td>
<td>-</td>
</tr>
<tr>
<td>EAP (RFC 3748)</td>
<td>o</td>
<td>-</td>
<td>-</td>
<td>o</td>
</tr>
<tr>
<td>IMAP (RFC 4314)</td>
<td>o</td>
<td>-</td>
<td>-</td>
<td>o</td>
</tr>
<tr>
<td>LDAP (RFC 4518)</td>
<td>o</td>
<td>-</td>
<td>o</td>
<td>o</td>
</tr>
</tbody>
</table>

Appendix B. The reason why local case mapping is alternative to case mapping in PRECIS framework

Local case mapping is alternative to Unicode Default Case Folding instead of being applied sequentially. Because, one outstanding issue regarding full case folding for characters is, some lowercase characters like "LATIN SMALL LETTER SHARP S" (U+00DF) (hereinafter referred to as "eszett") and ligatures like "LATIN SMALL LIGATURE FF" (U+FB00) that described in section Unconditional mappings of SpecialCasing.txt become a different codepoint by performing the case mapping using Unicode Default Case Folding in the PRECIS framework. In particular, German’s eszett can not keep the locale because eszett becomes two "LATIN SMALL LETTER S"s (U+0073 U+0073) by performing the case mapping using Unicode Default Case Folding. On the other hand, eszett doesn’t become a different codepoint by performing the case
mapping in SpecialCasing.txt. Therefore, if it is necessary to keep locale of characters, PRECIS profile designers should select local case mapping as alternative to Unicode Default Case Folding.

Appendix C. Limitation to local case mapping

As described in the section Section 2.3, the possible target characters of local case mapping are specified in SpecialCasing.txt. The Unicode Standard (at least, up to version 7.0.0) does not define any context-dependent mappings between "GREEK SMALL LETTER SIGMA" (U+03C3) (hereinafter referred to as "small sigma") and "GREEK SMALL LETTER FINAL SIGMA" (U+03C2) (hereinafter referred to as "final sigma"). Thus, local case mapping is not applicable to small sigma or final sigma, so case mapping in the PRECIS framework always maps final sigma to small sigma, independent of context, as also specified by Unicode Default Case Folding. (Note: Following comments are from SpecialCasing.txt.)

# Note: the following cases are not included, since they would case-fold in lowercasing
# 03C3; 03C2; 03A3; 03A3; Final_Sigma; # GREEK SMALL LETTER SIGMA
# 03C2; 03C3; 03A3; 03A3; Not_Final_Sigma; # GREEK SMALL LETTER

Appendix D. Change Log

D.1. Changes since -00

- Modify the Section 4.3 "Local case mapping" to specify the method to calculate codepoints that local case mapping targets.
- Add the Section 6 "Open issues".
- Modify the Section 7 "IANA Considerations".
- Modify the Section 8 "Security Considerations".
- Remove the "The initial PRECIS local case mapping registrations".
- Add the Appendix C "Code points list for local case mapping".
- Add the Appendix D "Change Log".

D.2. Changes since -01

- Unified PRECIS notation in all capital letters as well as other documents.
o Removed the Section 1 "Types of mapping" and the Section 2 "Protocol independent mapping" because width mapping is now in framework document.

o Added relationship between the framework document and this document in the Section 3 "Order of operations".

o Updated the Section 4 "Open issues" to address new issue raised on mailing list.

o Move the Section 6 "IANA Considerations" after the Section 5 "Security Considerations".

o Remove the Appendix B "Codepoints which need special mapping" and mentioned related documents in the Section 2.2.

D.3. Changes since -02

o Removed the "Open issues".

D.4. Changes since -03

o Modify the Section 1 "Introduction" in more clear text.

o Modify the Section 2.3 "Local case mapping" to clarify the purpose of the local case mapping and an example, and add restriction to use with PRECIS framework.

o Change the format in the Appendix B "Code points list for local case mapping".

o Split the Section 7 "References" into "Normative References" and "Informative References".

o Update the Unicode version 6.2 to 6.3 in this document.

D.5. Changes since -04

o Correct a sentence in the Section 2.3 "Local case mapping".

D.6. Changes since -05

o Correct some sentences in this document.

o Modify the local case mapping’s rule and target characters in the Section 2.3 "Local case mapping". This is to avoid user’s confusion towards Greek’s final sigma and German’s eszett.
Add the Section 4 "Open issues".

Modify the Section 8 "Security Considerations".

Modify the table format in the Appendix A. "Mapping type list each protocol".

Removed the Appendix B "Code points list for local case mapping".

Add the Appendix B "Local case mapping vs Case mapping".

D.7. Changes since -06

Removed the Section 4 "Open issues".

Change the title of the Appendix B "Local case mapping vs Case mapping" to "The reason why local case mapping is alternative to case mapping in PRECIS framework".

Add the Appendix C "Limitation to local case mapping".

D.8. Changes since -07

Modify the Section 1 "Introduction".

Modify the local case mapping’s rule and target characters in the Section 2.3 "Local case mapping".

Modify the Section 3 "Order of operations".

D.9. Changes since -08

Updated the Unicode version 6.3 to 7.0 in this document.

D.10. Changes since -09

Modify the Section 1 "Introduction" to clarify to the discussion of string matching and the use of mappings from the SpecialCasing.txt.

Modify the Section 2.3 "Local case mapping" to clarify to the discussion of string matching and the use of mappings from the SpecialCasing.txt.

Modify the Appendix B "The reason why local case mapping is alternative to case mapping in PRECIS framework" to state the result of the case mapping in SpecialCasing.txt of eszett.
D.11. Changes since -10

- Clarify the Appendix C "Limitation to local case mapping".

D.12. Changes since -11

- Correct a few sentence in the Section 2.3 "Local case mapping" to address comments by the IESG review.

- Removed citation part which includes "RECOMMENDED" (RFC 2119 word) in the Section 2.3 "Local case mapping" to avoid readers’ confusion.

- Modify the Section 4 "Security Considerations" to add a reference to RFC7564.
Authors’ Addresses

Yoshiro YONEYA
JPRS
Chiyoda First Bldg. East 13F
3-8-1 Nishi-Kanda
Chiyoda-ku, Tokyo  101-0065
Japan
Phone: +81 3 5215 8451
Email: yoshiro.yoneya@jprs.co.jp

Takahiro Nemoto
Keio University
Graduate School of Media Design
4-1-1 Hiyoshi, Kohoku-ku
Yokohama, Kanagawa  223-8526
Japan
Phone: +81 45 564 2517
Email: t.nemo10@kmd.keio.ac.jp