Internet Engineering Task Force (IETF)
Request for Comments: 7790
Category: Informational
ISSN: 2070-1721

Y. Yoneya
JPRS
T. Nemoto
Keio University

February 2016

Mapping Characters for Classes of the Preparation, Enforcement, and Comparison of Internationalized Strings (PRECIS)

Abstract

The framework for the 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 is 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 document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7790.

Table of Contents

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 cases; 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 ':' and '/' 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, the Extensible Authentication Protocol (EAP) [RFC3748], IMAP4 Access Control List (ACL) [RFC4314], and LDAPprep [RFC4518] define the rule that some code points 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. See also the "Note" in Section 2.1.1 of RFC 5646 [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 correspond to demands from applications that support users' locale and/or context. The complete set of possible target characters for local case mapping are the characters specified

in SpecialCasing.txt [Specialcasing] in 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 SpecialCasing.txt.

The case-folding method for a selected target character is to map into lowercase as defined in SpecialCasing.txt. The case-folding method for all other, non-target characters is as specified in 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 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 Section 5.2.2 of the PRECIS framework. Although the delimiter 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 for other users. Such confusion is not covered in any depth in this document.

5. References

5.1. Normative References

[RFC7564] Saint-Andre, P. and M. Blanchet, "PRECIS Framework: Preparation, Enforcement, and Comparison of Internationalized Strings in Application Protocols", RFC 7564, DOI 10.17487/RFC7564, May 2015, <http://www.rfc-editor.org/info/rfc7564>.

[Unicode] The Unicode Consortium, "The Unicode Standard, Version 7.0.0", (Mountain View, CA: The Unicode Consortium, 2014. ISBN 978-1-936213-09-2), <http://www.unicode.org/versions/Unicode7.0.0/>.

[Casefolding] The Unicode Consortium, "CaseFolding-7.0.0.txt", Unicode Character Database, July 2011, <http://www.unicode.org/Public/7.0.0/ucd/CaseFolding.txt>.

[Specialcasing] The Unicode Consortium, "SpecialCasing-7.0.0.txt", Unicode Character Database, July 2011, <http://www.unicode.org/Public/7.0.0/ucd/SpecialCasing.txt>.

5.2. Informative References

[RFC3454] Hoffman, P. and M. Blanchet, "Preparation of Internationalized Strings ("stringprep")", RFC 3454, DOI 10.17487/RFC3454, December 2002, <http://www.rfc-editor.org/info/rfc3454>.

[RFC3490] Faltstrom, P., Hoffman, P., and A. Costello, "Internationalizing Domain Names in Applications (IDNA)", RFC 3490, DOI 10.17487/RFC3490, March 2003, <http://www.rfc-editor.org/info/rfc3490>.

[RFC3491] Hoffman, P. and M. Blanchet, "Nameprep: A Stringprep Profile for Internationalized Domain Names (IDN)", RFC 3491, DOI 10.17487/RFC3491, March 2003, <http://www.rfc-editor.org/info/rfc3491>.

[RFC3722] Bakke, M., "String Profile for Internet Small Computer Systems Interface (iSCSI) Names", RFC 3722, DOI 10.17487/RFC3722, April 2004, <http://www.rfc-editor.org/info/rfc3722>.

[RFC3748] Aboba, B., Blunk, L., Vollbrecht, J., Carlson, J., and H. Levkowetz, Ed., "Extensible Authentication Protocol (EAP)", RFC 3748, DOI 10.17487/RFC3748, June 2004, <http://www.rfc-editor.org/info/rfc3748>.

[RFC4314] Melnikov, A., "IMAP4 Access Control List (ACL) Extension", RFC 4314, DOI 10.17487/RFC4314, December 2005, <http://www.rfc-editor.org/info/rfc4314>.

[RFC4518] Zeilenga, K., "Lightweight Directory Access Protocol (LDAP): Internationalized String Preparation", RFC 4518, DOI 10.17487/RFC4518, June 2006, <http://www.rfc-editor.org/info/rfc4518>.

[RFC5646] Phillips, A., Ed. and M. Davis, Ed., "Tags for Identifying Languages", BCP 47, RFC 5646, DOI 10.17487/RFC5646, September 2009, <http://www.rfc-editor.org/info/rfc5646>.

[RFC5895] Resnick, P. and P. Hoffman, "Mapping Characters for Internationalized Domain Names in Applications (IDNA) 2008", RFC 5895, DOI 10.17487/RFC5895, September 2010, <http://www.rfc-editor.org/info/rfc5895>.

Appendix A. Mapping Type List

A.1. Mapping Type List for Each Protocol

This table is the mapping type list for each protocol that uses the Stringprep framework [RFC3454] and is a PRECIS framework customer candidate (as Stringprep and the related IDNA versions in the table below are now obsolete). Values marked "o" indicate that the protocol uses the type of mapping. Values marked "-" indicate that the protocol doesn't use the type of mapping.

Protocol and mapping RFC	Width (NFKC)	Delimiter	Case	Special
IDNA [RFC3490]	-	o	-	-
IDNA [RFC3491]	o	-	o	-
iSCSI [RFC3722]	o	-	o	-
EAP [RFC3748]	o	-	-	o
IMAP [RFC4314]	o	-	-	o
LDAP [RFC4518]	o	-	o	o

Appendix B. Why Local Case Mapping Is an Alternative to Case Mapping in the PRECIS Framework

Local case mapping and Unicode Default Case Folding are alternatives. They can't be applied simultaneously or sequentially. One outstanding issue regarding full case folding for characters is that 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 are described in the "Unconditional mappings" section of SpecialCasing.txt become a different code point when the case mapping is performed using Unicode Default Case Folding in the PRECIS framework.

In particular, German's eszett cannot keep the locale because eszett becomes two "LATIN SMALL LETTER S"s (U+0073 U+0073) when the case mapping is performed using Unicode Default Case Folding. (See also 00DF in CaseFolding.txt [Casefolding].) On the other hand, eszett doesn't become a different code point when performing the case mapping in SpecialCasing.txt. Therefore, if it is necessary to keep the locale of characters, PRECIS profile designers should select local case mapping as an alternative to Unicode Default Case Folding.

Appendix C. Limitations of Local Case Mapping

As described in 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. The following comments are from SpecialCasing.txt. (Line breaks have been added due to line-length limitations.)

```
# 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 FINAL
  SIGMA
```

Acknowledgments

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

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 suggestions for this document during working group discussions.

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