Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 6, 2013

Y. YONEYA
JPRS
T. NEMOTO
Keio University

October 3, 2012

Mapping characters for precis classes
draft-yoneya-precis-mappings-03

Abstract

Preparation and comparison of internationalized strings ("precis") framework [I-D.ietf-precis-framework] is defining several classes of strings for preparation and comparison. In the document, case mapping is defined because many of protocols handle case sensitive or case insensitive string comparison and therefore preparation of string is mandatory. As described in IDNA mapping [RFC5895] and precis problem statement [I-D.ietf-precis-problem-statement], mappings in internationalized strings are not limited to case, but also width, delimiters and/or other specials are taken into consideration. This document is a guideline for authors of protocol profiles of precis framework and describes the mappings that must be considered between receiving user input and passing permitted code points to internationalized protocols.

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 April 6, 2013.

Copyright Notice

1. Introduction

In many cases, user input of internationalized strings is generated by input method editor ("IME") or copy-and-paste from free text. Usually users do not care case and/or width of input characters because they are identical for users' eyes. Further, users rarely switch IME state to input special characters such as protocol elements. For Internationalized Domain Names ("IDNs"), IDNA Mapping [RFC5895] describes methods to treat these issues. For precis strings, case mapping is defined as a process in precis framework [I-D.ietf-precis-framework], but width mapping, delimiter mapping and/or special mapping are not defined. Handling of mappings other than case is also important to increase chance of strings match as users expect. This document is a guideline for authors of protocol profiles of precis framework and describes the mappings that must be considered between receiving user input and passing permitted code points to internationalized protocols.

2. Types of mapping

This document defines two types of mapping. One is protocol independent mapping that doesn't depend on protocol rules and the other is protocol dependent mapping that depend on protocol rules. This document defines some mappings in these mapping types. Authors of protocol profiles of precis framework should need to give careful consideration to choice of mappings.

Each mapping type is described in following sections.

3. Protocol independent mapping

Protocol independent mapping is a mapping that doesn't depend on protocol rules.

3.1. Width mapping

Fullwidth and halfwidth characters (those defined with Decomposition Types <wide> and <narrow>) are mapped to their decomposition mappings as shown in the Unicode character database [Unicode].

Width mapping will increase backward compatibility with Stringprep [RFC3454] and precis framework [I-D.ietf-precis-framework]. Because in a Stringprep profile which specifies Unicode normalization form KC (NFKC) for normalization method, fullwidth/halfwidth characters are mapped into its compatible form. If a precis framework profile specified NFKC (which is not recommended), width mapping might not be useful.

4. Protocol dependent mapping

Protocol dependent mapping is a mapping that depend on protocol rules.

4.1. Delimiter mapping

Definitions of delimiters in certain protocols are differ from each other. Therefore, delimiter mapping table should be based on well defined mapping table for each protocol.

One of the most useful case of delimiter mapping is when FULL STOP character (U+002E) is a delimiter as well as domain name. Some of IME generates FULL STOP compatible characters such as IDEOGRAPHIC FULL STOP (U+3002) when users type FULL STOP on the keyboard.

4.2. Special mapping

Certain protocols have characters which need to map different character from precis framework defined mapping rule other than delimiter characters. In this document, these mappings are named special mapping. They are differ from each protocol. Therefore, special mapping table should be based on well defined mapping table for each protocol. Examples of special mapping are following;

- White spaces are mapped to SPACE (U+0020)
- Some characters such as control characters are mapped to nothing (Deletion)

LDAPprep[RFC4518] defines the rule that some codepoints(Appendix B.4) are mapped to SPACE (U+0020).

4.3. Local case mapping

Local case mapping is case folding that depend on language context. For example, given there is upper case I in a user ID strings, you should care what's language context that this user ID depend on when this character is mapped into lower case character. And if this depends on Turkish, the character should be mapped into LATIN SMALL LETTER DOTLESS I (U+0131) as this character's lower case.

This document defines characters that need local case mapping based on the Specialcasing.txt [Specialcasing] in section 3.13 of The Unicode Standerd [Unicode] to solve such a problem. Local case mapping targets only characters that get two different results to perfom just casefolding that is defined in the Casefolding.txt [Casefolding] and perfom special casefolding that is defined in the

Specialcasing.txt then casefolding, because precis framework have casefolding.

There are two types casefoldings defined as Unconditional Mappings and Conditional Mappings in the Specialcasing.txt. Conditional mappings have Language-Insensitive Mappings that targets characters whose full case mappings do not depend on language, but do depend on context and Language-Sensitive Mappings that these are characters whose full case mappings depend on language and perhaps also context.

Of these mappings, characters that Unconditional Mappings and Language-Insensitive Mappings in Conditional Mappings target are mapped into same codepoint(s) with just casefolding and special casefolding then casefolding. But characters that Language-Sensitive Mappings in Conditional Mappings targets are mapped into different codepoint with them. Therefore this document defined characters that are a part of characters of Lithuanian(lt), Turkish(tr) and Azerbaijanian(az) that Language-Sensitive Mappings targets as targets for local case mapping.

A list of characters that need Local case mapping are as follows.

Format:
<Language>; <Codepoint>; <Lowercase>; <Comments>

lt; 0049; 0069 0307; LATIN CAPITAL LETTER I
lt; 004A; 006A 0307; LATIN CAPITAL LETTER J
lt; 012E; 012F 0307; LATIN CAPITAL LETTER I WITH OGONEK
lt; 00CC; 0069 0307 0300; LATIN CAPITAL LETTER I WITH GRAVE
lt; 00CD; 0069 0307 0301; LATIN CAPITAL LETTER I WITH ACUTE
lt; 0128; 0069 0307 0303; LATIN CAPITAL LETTER I WITH TILDE
tr; 0130; 0069; LATIN CAPITAL LETTER I WITH DOT ABOVE
tr; 0049; 0131; LATIN CAPITAL LETTER I
az; 0130; 0069; LATIN CAPITAL LETTER I WITH DOT ABOVE
az; 0049; 0131; LATIN CAPITAL LETTER I

Section 6 "IANA Considerations" contains a template to registry these characters to IANA as precis local case mapping registry.

5. Applying order of mapping

Basically, applying order of mapping that this document describes aren't sensitive. This section defines applying order of mapping to minimize effect of codepoint change by mappings. This mapping order is very general and was designed to be acceptable to the widest user community.

1. width mapping

2. delimiter mapping

3. special mapping

4. local case mapping

5. precis framework

Mappings that this document describes should be performed before precis framework.

6. IANA Considerations

6.1. precis local case mapping registry

IANA is requested to create a registry of precis local case mapping. In accordance with [RFC5226], the registration policy is "RFC Required".

6.2. Template for precis local case mapping registry

The following information is to be given when a new precis local case mapping rule is created. The registration template is as follows:

Language: language name

Codepoint: Local case mapping that can be applied when this code point exists in the strings

Local lowercase: The lowercase codepoint after performing local case mapping

Comment: Character name of the code point

Appendix C contains further discussion and a table from which that registry can be initialized.

7. Security Considerations

TBD.

8. Acknowledgment

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

Pete Resnick et al. gave important suggestion for this document during at WG meeting.

9. References

[RFC3454] Hoffman, P. and M. Blanchet, "Preparation of Internationalized Strings ("stringprep")", RFC 3454, December 2002.

[RFC3490] Faltstrom, P., Hoffman, P., and A. Costello, "Internationalizing Domain Names in Applications (IDNA)", RFC 3490, March 2003.

[RFC3491] Hoffman, P. and M. Blanchet, "Nameprep: A Stringprep Profile for Internationalized Domain Names (IDN)", RFC 3491, March 2003.

[RFC3722] Bakke, M., "String Profile for Internet Small Computer Systems Interface (iSCSI) Names", RFC 3722, April 2004.

[RFC3748] Aboba, B., Blunk, L., Vollbrecht, J., Carlson, J., and H. Levkowetz, "Extensible Authentication Protocol (EAP)", RFC 3748, June 2004.

[RFC4013] Zeilenga, K., "SASLprep: Stringprep Profile for User Names and Passwords", RFC 4013, February 2005.

[RFC4314] Melnikov, A., "IMAP4 Access Control List (ACL) Extension", RFC 4314, December 2005.

[RFC4518] Zeilenga, K., "Lightweight Directory Access Protocol (LDAP): Internationalized String Preparation", RFC 4518, June 2006.

[RFC5895] Resnick, P. and P. Hoffman, "Mapping Characters for Internationalized Domain Names in Applications (IDNA) 2008", RFC 5895, September 2010.

[RFC6122] Saint-Andre, P., "Extensible Messaging and Presence Protocol (XMPP): Address Format", RFC 6122, March 2011.

[I-D.ietf-precis-framework] Saint-Andre, P. and M. Blanchet, "PRECIS Framework: Preparation and Comparison of Internationalized Strings in Application Protocols", draft-ietf-precis-framework-03 (work in progress), May 2012.

[I-D.ietf-precis-problem-statement] Blanchet, M. and A. Sullivan, "Stringprep Revision and PRECIS Problem Statement", draft-ietf-precis-problem-statement-06 (work in progress),

July 2012.

[Unicode] The Unicode Consortium, "The Unicode Standard, Version 6.1.0", <http://www.unicode.org/versions/Unicode6.1.0/>, 2012.

[Casefolding] "CaseFolding-6.1.0.txt", Unicode Character Database, July 2011, <http://www.unicode.org/Public/6.1.0/ucd/CaseFolding.txt>.

[Specialcasing] "SpecialCasing-6.1.0.txt", Unicode Character Database, July 2011, <http://www.unicode.org/Public/6.1.0/ucd/SpecialCasing.txt>.

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.

RFC \ Type of mapping	Width (NFKC)	Delimiter	Case	Special
3490	-	o	-	-
3491	o	-	o	-
3722	o	-	o	-
3748	o	-	-	o
4013	o	-	-	o
4314	o	-	-	o
4518	o	-	o	o
6120	-	-	o	-

Appendix B. Codepoints which need special mapping

B.1. RFC3748

Non-ASCII space characters [StringPrep, C.1.2] that can be mapped to SPACE (U+0020).

B.2. RFC4013

Non-ASCII space characters [StringPrep, C.1.2] that can be mapped to SPACE (U+0020).

B.3. RFC4314

Non-ASCII space characters [StringPrep, C.1.2] that can be mapped to SPACE (U+0020).

B.4. RFC4518

Codepoints mapped to SPACE (U+0020) are following;

U+0009 (CHARACTER TABULATION)
U+000A (LINE FEED (LF))
U+000B (LINE TABULATION)
U+000C (FORM FEED (FF))
U+000D (CARRIAGE RETURN (CR))
U+0085 (NEXT LINE (NEL))
U+0020 (SPACE)
U+00A0 (NO-BREAK SPACE)
U+1680 (OGHAM SPACE MARK)
U+2000 (EN QUAD)
U+2001 (EM QUAD)
U+2002 (EN SPACE)
U+2003 (EM SPACE)
U+2004 (THREE-PER-EM SPACE)
U+2005 (FOUR-PER-EM SPACE)
U+2006 (SIX-PER-EM SPACE)
U+2007 (FIGURE SPACE)
U+2008 (PUNCTUATION SPACE)
U+2009 (THIN SPACE)
U+200A (HAIR SPACE)
U+2028 (Line Separator)
U+2029 (Paragraph Separator)
U+202F (NARROW NO-BREAK SPACE)
U+205F (MEDIUM MATHEMATICAL SPACE)
U+3000 (IDEOGRAPHIC SPACE)

All other control code (e.g., Cc) points or code points with a

control function (e.g., Cf) are mapped to nothing. Codepoints mapped to nothing that aren't specified by Stringprep are following;

U+0000-0008
U+000E-001F
U+007F-0084
U+0086-009F
U+06DD
U+070F
U+180E
U+200E-200F
U+202A-202E
U+2061-2063
U+206A-206F
U+FFF9-FFFB
U+1D173-1D17A
U+E0001
U+E0020-E007F

Appendix C. The initial precis local case mapping registrations

C.1. Lithuanian

language: Lithuanian

Codepoint: U+0049
Local lowercase: U+0069 U+0307
Comment: LATIN CAPITAL LETTER I

Codepoint: U+004A
Local lowercase: U+006A U+0307
Comment: LATIN CAPITAL LETTER J

Codepoint: U+012E
Local lowercase: U+012F U+0307
Comment: LATIN CAPITAL LETTER I WITH OGONEK

Codepoint: U+00CC
Local lowercase: U+0069 U+0307 U+0300
Comment: LATIN CAPITAL LETTER I WITH GRAVE

Codepoint: U+00CD
Local lowercase: U+0069 U+0307 U+0301
Comment: LATIN CAPITAL LETTER I WITH ACUTE

Codepoint: U+0128
Local lowercase: U+0069 U+0307 U+0303
Comment: LATIN CAPITAL LETTER I WITH TILDE

C.2. Turkish

language: Turkish

Codepoint: U+0130
Local lowercase: U+0069
Comment: LATIN CAPITAL LETTER I WITH DOT ABOVE

Codepoint: U+0049
Local lowercase: U+0131
Comment: LATIN CAPITAL LETTER I

C.3. Azerbaijanian

language: Azerbaijanian

Codepoint: U+0130
Local lowercase: U+0069

Comment: LATIN CAPITAL LETTER I WITH DOT ABOVE

Codepoint: U+0049
Local lowercase: U+0131
Comment: LATIN CAPITAL LETTER I

Appendix D. Change Log

D.1. Changes since -00

- Add the Section 2.3 "Special mapping" in Section 2 Type of mappings.
- Add the topic about the special mapping and additional case mapping in Section 3 "Discussion".
- Add Appendices;
  Appendix A "Mapping type list each protocols"
  Appendix B "Code point list is need special mapping"
  Appendix D "Change Log"
- Add the Section 8 "Acknowledgment".

D.2. Changes since -01

- Modify document structure as a guideline for authors of protocol profiles of precis framework.
- Group mappings that this document defines into two types.
- Add the Section 5 "Applying order of mapping".
- Delete the section 3 "Discussion".

D.3. Changes since -02

- Modify the Section 4.3 "Local case mapping" for defining characters that local case mapping targets.
- Request creating registry of precis local case mapping to IANA and define a template for registry of precis local case mapping in the Section 6 "IANA Considerations".
- Add the Appendix C "The initial precis local case mapping registrations".

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