Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 10, 2019

P. Faltstrom
Netnod

October 07, 2018

IDNA2008 and Unicode 11.0.0
draft-faltstrom-unicode11-04

Abstract

This document describes changes between Unicode 6.3.0 and Unicode 11.0.0 in the context of IDNA2008. It further suggests for the IETF a path forward regarding ensuring IDNA2008 follows the evolution of the Unicode Standard.

In a few cases changes have been made in the Unicode Standard related to the algorithm IDNA2008 specifies. IDNA2008 do give the ability to add exceptions for backward compatibility to the algorithm but the conclusions provided in this document suggests no such changes.

Thus this document requests that IANA update the tables to Unicode 11.

In addition, all registries should continue the practice of calculating a repertoire using conservatism and inclusion principles.

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 https://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 10, 2019.

Copyright Notice

Copyright (c) 2018 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction

The current version of Internationalized Domain Names for Applications (IDNA) was largely completed in 2008, known within the series and elsewhere as "IDNA2008" and is specified in a series of documents (see Section Section 3.1). The standard include an

algorithm by which a derived property value is calculated based on the properties defined in the Unicode Standard.

When the Unicode Standard is updated code points are assigned and property values might be changed for already assigned code points.

Assigning code points might create problems if the newly assigned code points are compositions of code points so that it either changes or would have changed the normalization functions. This because it changes the matching algorithms used which in turn might create problems looking up already stored strings in for example DNS.

Changing properties for already assigned code points might create problems if the change do result in the derived property value changes. This might make an earlier allowed code point (derived property value PVALID) not be allowed anymore (derived property value DISALLOWED). Or the other way around, a code point that was not allowed (and because of that blocked in some situations) suddenly end up being allowed.

Historically the IETF has accepted all implications of changes in the Unicode Standard even though the changes have resulted in problematic changes in the derived property value. The primary reason for that is that staying with the Unicode Standard has been viewed as important given the diversity in implementations already existing in the wild.

As described in Section 4, a few changes have been made regarding certain attributes to code points in Unicode between version 6.3.0 and 11.0.0. Such changes could result in either a change in the derived property value for the code point in question or no such change. In turn, if the result is a change, it can be between any of the derived property values except DISALLOWED. Also in this case, when moving from version 6.3.0 to 11.0.0, this document concludes that no exceptions are to be added to IDNA2008 even if changes in the derived property value is a result of the changes made in Unicode.

Specifically, the Internet Architecture Board did issue a statement [IAB] which requested IETF to resolve the issues related to the code point ARABIC LETTER BEH WITH HAMZA ABOVE (U+08A1), introduced in Unicode 7.0.0 [Unicode-7.0.0]. This document resolves this issue and suggests IDNA2008 standard is to follow the Unicode Standard and not update RFC 5892 [RFC5892] or any other IDNA2008 RFCs.

2. Keywords for Requirement Levels

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

3. Background

3.1. IDNA2008 Documents

IDNA2008 consists of the following documents:

- A document, RFC 5890 [RFC5890], containing definitions and other material that are needed for understanding other documents in the set. It is referred to informally in other documents in the set as "Defs" or "Definitions".

- A document, RFC 5891 [RFC5891], that describes the core IDNA2008 protocol and its operations. It is to be interpreted in combination with the Bidi document, described immediately below. It is referred to informally in other documents in the set as "Protocol".

- A specification, RFC 5892 [RFC5892], of the categories and rules that identify the code points allowed in a label written in native character form (defined more specifically as a "U-label"), based originally on Unicode 5.2.0 [Unicode-5.2.0] code point assignments and additional rules unique to IDNA2008. The Unicode-based rules are expected to be stable across Unicode updates and hence independent of Unicode versions. That specification obsoletes RFC 3491 [RFC3491] and IDN use of the tables to which it refers. It is referred to informally in other documents in the set as "Tables".

- A document, RFC 5893 [RFC5893], that specifies special rules (Bidi) for labels that contain characters that are written from right to left.

- A document, RFC 5894 [RFC5894], that provides an overview of the protocol and associated tables together with explanatory material and some rationale for the decisions that led to IDNA2008. That document also contains advice for registry operations and those who use Internationalized Domain Names (IDNs). It is referred to informally in other documents in the set as "Rationale".

- A document, RFC 5895 [RFC5895], that discusses the issue of mapping characters into other characters and that provides guidance for doing so when that is appropriate. That document,

referred to informally as "Mapping", provides advice; it is not a required part of IDNA.

- A document, RFC 6452 [RFC6452], that looks at some changes made to Unicode 6.0.0 [Unicode-6.0.0] that resulted in the derived property value change for the code points U+0CF1, U+0CF2 and U+19DA. The first two changed from DISALLOWED to PVALID, the last from PVALID to DISSALOWED. IETF came to the conclusion the changes where acceptable and RFC 5892 [RFC5892] was not updated to make the derived property value not change for these code points.

3.2. Deployment

The deployment of IDNA2008 is unfortunately quite diverse. The following lists some of the strategies that existing implementations are known to implement:

- IDNA2003 as specified in RFC 3490 [RFC3490] and RFC 3491 [RFC3491] which implies using a table within which it is said whether code points are allowed to be used or not, and this after doing the in IDNA2003 included normalization.

- A mix between IDNA2003 and IDNA2008 where code points assigned to Unicode after Unicode 3.2.0 [Unicode-3.2.0] have derived property value calculated according to the algorithm specified in IDNA2008.

- Strict IDNA2008 following IANA which implies stayed at Unicode 6.3.0 [Unicode-6.3.0] and treating later assigned code points as UNASSIGNED.

- The IDNA2008 algorithm applied to whatever version of Unicode Standard exists in the operating system and/or libraries used, regardless of whether the version is later than Unicode version 6.3.0 or not.

- A mix between IDNA2003 and IDNA2008 according to local interpretation of the Unicode Technical Standard #46 [UTS-46].

The issue is further complicated by having a very diverse implementations of the requirements in RFC 5894 [RFC5894] that registry operators to based on the IDNA2008 specification create additional rules for what code points are allowed to be used for registration.

In practice, the Unicode Consortium creates a maximum set of code points by assigning code points in the Unicode Standard. The IDNA2008 rules based on the Unicode Standard create a subset of these by assigning the PVALID derived property value to them. Registries

(and others dealing with Internationalized Domain Names) are supposed to create an even smaller subset that ultimately is the set of code points that can be used in a particular registry.

There is further recommendation to be conservative when these subsets are calculated and to use the inclusion principle; this is explained in SAC-084 [SAC-084] and RFC 6912 [RFC6912].

The complicated situation with deployment of IDNA2008 is discussed further in draft-klensin-idna-rfc5891bis [I-D.klensin-idna-rfc5891bis] and draft-freytag-troublesome-characters [I-D.freytag-troublesome-characters].

4. Notable changes between Unicode 6.3.0 and 11.0.0

4.1. Changes to Unicode 7.0.0

The character ARABIC LETTER BEH WITH HAMZA ABOVE U+08A1 was introduced in Unicode 7.0.0. This was discussed in the IETF extensively and by IAB in their statement [IAB] requesting the IETF to investigate the issue. Specifically IAB stated:

> On the same precautionary principle, the IAB recommends that the Internationalized Domain Names for Applications (IDNA) Parameters registry (http://www.iana.org/assignments/idna-tables/) not be updated to Unicode 7.0.0 until the IETF has consensus on a solution to this problem.

The discussion in the IETF concluded that although it is possible to create "the same" character in multiple ways, the issue with U+08A1 is not unique. In the case of U+08A1, it can be represented with the sequence ARABIC LETTER BEH (U+0628) and ARABIC HAMZA ABOVE (U+0654). Just like LATIN SMALL LETTER A WITH DIAERESIS (U+00E4) can be represented via the sequence LATIN SMALL LETTER A (U+0061), and COMBINING DIAERESIS (U+0308). One difference between these sequences is how they are treated in the normalization forms specified by the Unicode Consortium.

As U+08A1 is discussed in draft-freytag-troublesome-characters [I-D.freytag-troublesome-characters] and elsewhere. Regardless of whether those discussions ends in recommending including the code point in the repertoire of characters permissable for registration or not, it is acceptable to allow the code point to have a derived property value of PVALID.

4.2. Changes between Unicode 7.0.0 and 10.0.0

There are no changes made to Unicode between version 7.0.0 and 10.0.0 that impact IDNA2008 calculation of the derived property value.

4.3. Changes to Unicode 11.0.0

The Unicode Standard Version 11.0.0 [Unicode-11.0.0] has included a number of changes [Changes-11.0.0] from version 10.0.0, specifically to UnicodeData.txt:

- Entries were added for the 684 new characters, including letters, combining marks, digits, symbols, and punctuation marks.
- Georgian letters in the ranges U+10D0..U+10FA, U+10FD..U+10FF were changed from Lo to Ll, to reflect their status as the lowercase of new Georgian case pairs. Case mappings were also added.
- U+111C9 SHARADA SANDHI MARK was changed from Po to Mn, and from bc=L to bc=NSM.
- U+11A07 ZANABAZAR SQUARE VOWEL SIGN AI and U+11A08 ZANABZAR SQUARE VOWEL SIGN AU were corrected from Mc to Mn.
- U+29A1 SPHERICAL ANGLE OPENING UP was changed to Bidi_M=N.

These changes to the Unicode Standard have the following implications for these code points:

- The newly assigned 684 characters are to have a derived property value as of a result of applying the IDNA2008 algorithm.
- The Georgian letters in the ranges U+10D0..U+10FA and U+10FD..U+10FF have existed since before IDNA2008 was created. Applying the IDNA2008 algorithm to the code points did assign the derived property value PVALID and that value is unchanged even if the underlying Unicode properties have changed.
- The U+111C9 SHARADA SANDHI MARK was added to Unicode 8.0.0 [Unicode-8.0.0]. Applying the IDNA2008 algorithm to the code point did assign the derived property value DISALLOWED. The changes in the underlying properties in the Unicode Standard Version 11.0.0 [Unicode-11.0.0] make the derived property value change to PVALID which is an acceptable change.
- The characters U+11A07 ZANABAZAR SQUARE VOWEL SIGN AI and U+11A08 ZANABZAR SQUARE VOWEL SIGN AU were added to Unicode 10.0.0 [Unicode-10.0.0]. Applying the IDNA2008 algorithm to the code

points did assign the derived property value PVALID and that value is unchanged even if the underlying Unicode properties have changed.

- U+29A1 SPHERICAL ANGLE OPENING UP have existed since before IDNA2008 was created. Applying the IDNA2008 algorithm to the code point did assign the derived property value PVALID and that value is unchanged even if the underlying Unicode properties have changed.

5. Conclusion

As described in Section 4 changes have been made to Unicode between version 6.3.0 and 11.0.0. Some changes to specific characters changed their derived property value. Others did not. Given the diverse deployment described in Section 3.2 and the changes described, including implications to normalization, the conclusion is to not add any exception rules to IDNA2008.

To increase overall harmonization in the use of internationalized domain names, the author recommends that the derived property values MUST be calculated according to the IDNA2008 specification for Unicode Version 11.0.0 [Unicode-11.0.0].

All registries (and others) SHOULD calculate a repertoire, for example as explained in draft-freytag-troublesome-characters [I-D.freytag-troublesome-characters] and draft-klensin-idna-rfc5891bis [I-D.klensin-idna-rfc5891bis] using the conservatism and inclusion principles as laid out in SAC-084 [SAC-084].

6. IANA Considerations

IANA is requested to update the registry of derived property values after validation with the Appointed Expert that the derived property values are calculated correctly.

7. Security Considerations

This document makes recommendations regarding the use of the IDNA2008 algorithm for calculation of derived property values, based on the current Unicode version. It also recommends that registries (and others dealing with Internationalized Domain Names) explicitly select appropriate subsets of characters with the derived value of PVALID.

Not following these recommendations can lead to various security issues. Specifically, allowing confusable characters may lead to various phishing attacks.

8. Acknowledgements

Thanks to Martin Durst, Asmus Freytag, Ted Hardie, John Klensin, Erik Nordmark, Michel Suignard, Andrew Sullivan and Suzanne Woolf for input to this document.

9. References

9.1. Normative References

[IAB] Internet Architecture Board, "IAB Statement on Identifiers and Unicode 7.0.0", IAB Statement on Identifiers and Unicode 7.0.0 https://www.iab.org/documents/correspondence-reports-documents/2015-2/iab-statement-on-identifiers-and-unicode-7-0-0/, January 2015.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

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

[RFC5890] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework", RFC 5890, DOI 10.17487/RFC5890, August 2010, <https://www.rfc-editor.org/info/rfc5890>.

[RFC5891] Klensin, J., "Internationalized Domain Names in Applications (IDNA): Protocol", RFC 5891, DOI 10.17487/RFC5891, August 2010, <https://www.rfc-editor.org/info/rfc5891>.

[RFC5892] Faltstrom, P., Ed., "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)", RFC 5892, DOI 10.17487/RFC5892, August 2010, <https://www.rfc-editor.org/info/rfc5892>.

[RFC5893] Alvestrand, H., Ed. and C. Karp, "Right-to-Left Scripts for Internationalized Domain Names for Applications (IDNA)", RFC 5893, DOI 10.17487/RFC5893, August 2010, <https://www.rfc-editor.org/info/rfc5893>.

[RFC6452] Faltstrom, P., Ed. and P. Hoffman, Ed., "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA) - Unicode 6.0", RFC 6452, DOI 10.17487/RFC6452, November 2011, <https://www.rfc-editor.org/info/rfc6452>.

9.2. Non-normative references

[Changes-11.0.0]
The Unicode Consortium, "Unicode Standard Annex #44", Unicode Standard Annex #44, UNICODE CHARACTER DATABASE, Change History https://www.unicode.org/reports/tr44/tr44-21d4.html#Change_History, May 2018.

[I-D.freytag-troublesome-characters]
Freytag, A., Klensin, J., and A. Sullivan, "Those Troublesome Characters: A Registry of Unicode Code Points Needing Special Consideration When Used in Network Identifiers", draft-freytag-troublesome-characters-01 (work in progress), June 2017.

[I-D.klensin-idna-rfc5891bis]
Klensin, J. and A. Freytag, "Internationalized Domain Names in Applications (IDNA): Registry Restrictions and Recommendations", draft-klensin-idna-rfc5891bis-01 (work in progress), September 2017.

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

[RFC5894] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Background, Explanation, and Rationale", RFC 5894, DOI 10.17487/RFC5894, August 2010, <https://www.rfc-editor.org/info/rfc5894>.

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

[RFC6912] Sullivan, A., Thaler, D., Klensin, J., and O. Kolkman, "Principles for Unicode Code Point Inclusion in Labels in the DNS", RFC 6912, DOI 10.17487/RFC6912, April 2013, <https://www.rfc-editor.org/info/rfc6912>.

[SAC-084] The Security and Stability Advisory Committee, "SAC084", SSAC Comments on Guidelines for the Extended Process Similarity Review Panel for the IDN ccTLD Fast Track Process https://www.icann.org/en/system/files/files/sac-084-en.pdf, August 2016.

[Unicode-10.0.0]
The Unicode Consortium, "The Unicode Standard, Version 10.0.0", The Unicode Standard, Version 10.0.0 ISBN 978-1-936213-16-0, June 2017.

[Unicode-11.0.0]
The Unicode Consortium, "The Unicode Standard, Version 11.0.0", The Unicode Standard, Version 11.0.0 ISBN 978-1-936213-19-1, June 2018.

[Unicode-3.2.0]
The Unicode Consortium, "The Unicode Standard, Version 3.2.0", The Unicode Standard, Version 3.2.0 ISBN 0-201-61633-5, March 2002.

[Unicode-5.2.0]
The Unicode Consortium, "The Unicode Standard, Version 5.2.0", The Unicode Standard, Version 5.2 ISBN 978-1-936213-00-9, October 2009.

[Unicode-6.0.0]
The Unicode Consortium, "The Unicode Standard, Version 6.0.0", The Unicode Standard, Version 6.0.0 ISBN 978-1-936213-01-6, October 2011.

[Unicode-6.3.0]
The Unicode Consortium, "The Unicode Standard, Version 6.3.0", The Unicode Standard, Version 6.3.0 ISBN 978-1-936213-08-5, September 2013.

[Unicode-7.0.0]
The Unicode Consortium, "The Unicode Standard, Version 7.0.0", The Unicode Standard, Version 7.0.0 ISBN 978-1-936213-09-2, June 2014.

[Unicode-8.0.0]
The Unicode Consortium, "The Unicode Standard, Version 8.0.0", The Unicode Standard, Version 8.0.0 ISBN 978-1-936213-10-8, June 2015.

[Unicode-9.0.0]
The Unicode Consortium, "The Unicode Standard, Version 9.0.0", The Unicode Standard, Version 9.0.0 ISBN 978-1-936213-13-9, June 2016.

[UTS-46] The Unicode Consortium, "Unicode Technical Standard #46, Version 11.0.0", UNICODE IDNA COMPATIBILITY PROCESSING http://www.unicode.org/reports/tr46/, May 2018.

Appendix A. Changes from Unicode 6.3.0 to Unicode 7.0.0

Changes from derived property value UNASSIGNED to either PVALID or DISALLOWED.

```
037F        ; DISALLOWED  # GREEK CAPITAL LETTER YOT
0528..052F  ; DISALLOWED  # CYRILLIC CAPITAL LETTER EN WITH LEFT HOOK..C
058D..058E  ; DISALLOWED  # RIGHT-FACING ARMENIAN ETERNITY SIGN..LEFT-FA
0605        ; DISALLOWED  # ARABIC NUMBER MARK ABOVE
08A1        ; PVALID      # ARABIC LETTER BEH WITH HAMZA ABOVE
08AD..08B2  ; PVALID      # ARABIC LETTER LOW ALEF..ARABIC LETTER ZAIN W
08FF        ; PVALID      # ARABIC MARK SIDEWAYS NOON GHUNNA
0978        ; PVALID      # DEVANAGARI LETTER MARWARI DDA
0980        ; PVALID      # BENGALI ANJI
0C00        ; PVALID      # TELUGU SIGN COMBINING CANDRABINDU ABOVE
0C34        ; PVALID      # TELUGU LETTER LLLA
0C81        ; PVALID      # KANNADA SIGN CANDRABINDU
0D01        ; PVALID      # MALAYALAM SIGN CANDRABINDU
0DE6..0DEF  ; PVALID      # SINHALA LITH DIGIT ZERO..SINHALA LITH DIGIT
16F1..16F8  ; PVALID      # RUNIC LETTER K..RUNIC LETTER FRANKS CASKET A
191D..191E  ; PVALID      # LIMBU LETTER GYAN..LIMBU LETTER TRA
1AB0..1ABE  ; PVALID      # COMBINING DOUBLED CIRCUMFLEX ACCENT..COMBINI
1CF8..1CF9  ; PVALID      # VEDIC TONE RING ABOVE..VEDIC TONE DOUBLE RIN
1DE7..1DF5  ; PVALID      # COMBINING LATIN SMALL LETTER ALPHA..COMBININ
20BB..20BD  ; DISALLOWED  # NORDIC MARK SIGN..RUBLE SIGN
23F4..23FA  ; DISALLOWED  # BLACK MEDIUM LEFT-POINTING TRIANGLE..BLACK C
2700        ; DISALLOWED  # BLACK SAFETY SCISSORS
2B4D..2B4F  ; DISALLOWED  # DOWNWARDS TRIANGLE-HEADED ZIGZAG ARROW..SHOR
2B5A..2B73  ; DISALLOWED  # SLANTED NORTH ARROW WITH HOOKED HEAD..DOWNWA
2B76..2B95  ; DISALLOWED  # NORTH WEST TRIANGLE-HEADED ARROW TO BAR..RIG
2B98..2BB9  ; DISALLOWED  # THREE-D TOP-LIGHTED LEFTWARDS EQUILATERAL AR
2BBD..2BC8  ; DISALLOWED  # BALLOT BOX WITH LIGHT X..BLACK MEDIUM RIGHT-
2BCA..2BD1  ; DISALLOWED  # TOP HALF BLACK CIRCLE..UNCERTAINTY SIGN
2E3C..2E42  ; DISALLOWED  # STENOGRAPHIC FULL STOP..DOUBLE LOW-REVERSED-
A698..A69D  ; DISALLOWED  # CYRILLIC CAPITAL LETTER DOUBLE O..MODIFIER L
A794..A79F  ; PVALID      # LATIN SMALL LETTER C WITH PALATAL HOOK..LATI
A7AB..A7AD  ; DISALLOWED  # LATIN CAPITAL LETTER REVERSED OPEN E..LATIN
A7B0..A7B1  ; DISALLOWED  # LATIN CAPITAL LETTER TURNED K..LATIN CAPITAL
A7F7        ; PVALID      # LATIN EPIGRAPHIC LETTER SIDEWAYS I
```

```
A9E0..A9FE  ; PVALID      # MYANMAR LETTER SHAN GHA..MYANMAR LETTER TAI
AA7C..AA7F  ; PVALID      # MYANMAR SIGN TAI LAING TONE-2..MYANMAR LETTE
AB30..AB5F  ; PVALID      # LATIN SMALL LETTER BARRED ALPHA..MODIFIER LE
AB64..AB65  ; PVALID      # LATIN SMALL LETTER INVERTED ALPHA..GREEK LET
FE27..FE2D  ; PVALID      # COMBINING LIGATURE LEFT HALF BELOW..COMBININ
1018B..1018C; DISALLOWED  # GREEK ONE QUARTER SIGN..GREEK SINUSOID SIGN
101A0       ; DISALLOWED  # GREEK SYMBOL TAU RHO
102E0..102FB; PVALID      # COPTIC EPACT THOUSANDS MARK..COPTIC EPACT NU
1031F       ; PVALID      # OLD ITALIC LETTER ESS
10350..1037A; PVALID      # OLD PERMIC LETTER AN..COMBINING OLD PERMIC L
10500..10527; PVALID      # ELBASAN LETTER A..ELBASAN LETTER KHE
10530..10563; PVALID      # CAUCASIAN ALBANIAN LETTER ALT..CAUCASIAN ALB
1056F       ; DISALLOWED  # CAUCASIAN ALBANIAN CITATION MARK
10600..10736; PVALID      # LINEAR A SIGN AB001..LINEAR A SIGN A664
10740..10755; PVALID      # LINEAR A SIGN A701 A..LINEAR A SIGN A732 JE
10760..10767; PVALID      # LINEAR A SIGN A800..LINEAR A SIGN A807
10860..1089E; PVALID      # PALMYRENE LETTER ALEPH..NABATAEAN LETTER TAW
108A7..108AF; DISALLOWED  # NABATAEAN NUMBER ONE..NABATAEAN NUMBER ONE H
10A80..10A9F; PVALID      # OLD NORTH ARABIAN LETTER HEH..OLD NORTH ARAB
10AC0..10AE6; PVALID      # MANICHAEAN LETTER ALEPH..MANICHAEAN ABBREVIA
10AEB..10AF6; DISALLOWED  # MANICHAEAN NUMBER ONE..MANICHAEAN PUNCTUATIO
10B80..10B91; PVALID      # PSALTER PAHLAVI LETTER ALEPH..PSALTER PAHLAV
10B99..10B9C; DISALLOWED  # PSALTER PAHLAVI SECTION MARK..PSALTER PAHLAV
10BA9..10BAF; DISALLOWED  # PSALTER PAHLAVI NUMBER ONE..PSALTER PAHLAVI
1107F       ; PVALID      # BRAHMI NUMBER JOINER
11150..11176; PVALID      # MAHAJANI LETTER A..MAHAJANI LIGATURE SHRI
111CD       ; DISALLOWED  # SHARADA SUTRA MARK
111DA       ; PVALID      # SHARADA EKAM
111E1..111F4; DISALLOWED  # SINHALA ARCHAIC DIGIT ONE..SINHALA ARCHAIC N
11200..11211; PVALID      # KHOJKI LETTER A..KHOJKI LETTER JJA
11213..1123D; PVALID      # KHOJKI LETTER NYA..KHOJKI ABBREVIATION SIGN
112B0..112EA; PVALID      # KHUDAWADI LETTER A..KHUDAWADI SIGN VIRAMA
112F0..112F9; PVALID      # KHUDAWADI DIGIT ZERO..KHUDAWADI DIGIT NINE
11301..11303; PVALID      # GRANTHA SIGN CANDRABINDU..GRANTHA SIGN VISAR
11305..1130C; PVALID      # GRANTHA LETTER A..GRANTHA LETTER VOCALIC L
1130F..11310; PVALID      # GRANTHA LETTER EE..GRANTHA LETTER AI
11313..11328; PVALID      # GRANTHA LETTER OO..GRANTHA LETTER NA
1132A..11330; PVALID      # GRANTHA LETTER PA..GRANTHA LETTER RA
11332..11333; PVALID      # GRANTHA LETTER LA..GRANTHA LETTER LLA
11335..11339; PVALID      # GRANTHA LETTER VA..GRANTHA LETTER HA
1133C..11344; PVALID      # GRANTHA SIGN NUKTA..GRANTHA VOWEL SIGN VOCAL
11347..11348; PVALID      # GRANTHA VOWEL SIGN EE..GRANTHA VOWEL SIGN AI
1134B..1134D; PVALID      # GRANTHA VOWEL SIGN OO..GRANTHA SIGN VIRAMA
11357       ; PVALID      # GRANTHA AU LENGTH MARK
1135D..11363; PVALID      # GRANTHA SIGN PLUTA..GRANTHA VOWEL SIGN VOCAL
11366..1136C; PVALID      # COMBINING GRANTHA DIGIT ZERO..COMBINING GRAN
11370..11374; PVALID      # COMBINING GRANTHA LETTER A..COMBINING GRANTH
11480..114C7; PVALID      # TIRHUTA ANJI..TIRHUTA OM
```

```
114D0..114D9; PVALID      # TIRHUTA DIGIT ZERO..TIRHUTA DIGIT NINE
11580..115B5; PVALID      # SIDDHAM LETTER A..SIDDHAM VOWEL SIGN VOCALIC
115B8..115C9; PVALID      # SIDDHAM VOWEL SIGN E..SIDDHAM END OF TEXT MA
11600..11644; PVALID      # MODI LETTER A..MODI SIGN HUVA
11650..11659; PVALID      # MODI DIGIT ZERO..MODI DIGIT NINE
118A0..118F2; DISALLOWED  # WARANG CITI CAPITAL LETTER NGAA..WARANG CITI
118FF       ; PVALID      # WARANG CITI OM
11AC0..11AF8; PVALID      # PAU CIN HAU LETTER PA..PAU CIN HAU GLOTTAL S
1236F..12398; PVALID      # CUNEIFORM SIGN KAP ELAMITE..CUNEIFORM SIGN U
12463..1246E; DISALLOWED  # CUNEIFORM NUMERIC SIGN ONE QUARTER GUR..CUNE
12474       ; DISALLOWED  # CUNEIFORM PUNCTUATION SIGN DIAGONAL QUADCOLO
16A40..16A5E; PVALID      # MRO LETTER TA..MRO LETTER TEK
16A60..16A69; PVALID      # MRO DIGIT ZERO..MRO DIGIT NINE
16A6E..16A6F; DISALLOWED  # MRO DANDA..MRO DOUBLE DANDA
16AD0..16AED; PVALID      # BASSA VAH LETTER ENNI..BASSA VAH LETTER I
16AF0..16AF5; PVALID      # BASSA VAH COMBINING HIGH TONE..BASSA VAH FUL
16B00..16B45; PVALID      # PAHAWH HMONG VOWEL KEEB..PAHAWH HMONG SIGN C
16B50..16B59; PVALID      # PAHAWH HMONG DIGIT ZERO..PAHAWH HMONG DIGIT
16B5B..16B61; DISALLOWED  # PAHAWH HMONG NUMBER TENS..PAHAWH HMONG NUMBE
16B63..16B77; PVALID      # PAHAWH HMONG SIGN VOS LUB..PAHAWH HMONG SIGN
16B7D..16B8F; PVALID      # PAHAWH HMONG CLAN SIGN TSHEEJ..PAHAWH HMONG
1BC00..1BC6A; PVALID      # DUPLOYAN LETTER H..DUPLOYAN LETTER VOCALIC M
1BC70..1BC7C; PVALID      # DUPLOYAN AFFIX LEFT HORIZONTAL SECANT..DUPLO
1BC80..1BC88; PVALID      # DUPLOYAN AFFIX HIGH ACUTE..DUPLOYAN AFFIX HI
1BC90..1BC99; PVALID      # DUPLOYAN AFFIX LOW ACUTE..DUPLOYAN AFFIX LOW
1BC9C..1BCA3; DISALLOWED  # DUPLOYAN SIGN O WITH CROSS..SHORTHAND FORMAT
1E800..1E8C4; PVALID      # MENDE KIKAKUI SYLLABLE M001 KI..MENDE KIKAKU
1E8C7..1E8D6; DISALLOWED  # MENDE KIKAKUI DIGIT ONE..MENDE KIKAKUI COMBI
1F0BF       ; DISALLOWED  # PLAYING CARD RED JOKER
1F0E0..1F0F5; DISALLOWED  # PLAYING CARD FOOL..PLAYING CARD TRUMP-21
1F10B..1F10C; DISALLOWED  # DINGBAT CIRCLED SANS-SERIF DIGIT ZERO..DINGB
1F321..1F32C; DISALLOWED  # THERMOMETER..WIND BLOWING FACE
1F336       ; DISALLOWED  # HOT PEPPER
1F37D       ; DISALLOWED  # FORK AND KNIFE WITH PLATE
1F394..1F39F; DISALLOWED  # HEART WITH TIP ON THE LEFT..ADMISSION TICKET
1F3C5       ; DISALLOWED  # SPORTS MEDAL
1F3CB..1F3CE; DISALLOWED  # WEIGHT LIFTER..RACING CAR
1F3D4..1F3DF; DISALLOWED  # SNOW CAPPED MOUNTAIN..STADIUM
1F3F1..1F3F7; DISALLOWED  # WHITE PENNANT..LABEL
1F43F       ; DISALLOWED  # CHIPMUNK
1F441       ; DISALLOWED  # EYE
1F4F8       ; DISALLOWED  # CAMERA WITH FLASH
1F4FD..1F4FE; DISALLOWED  # FILM PROJECTOR..PORTABLE STEREO
1F53E..1F53F; DISALLOWED  # LOWER RIGHT SHADOWED WHITE CIRCLE..UPPER RIG
1F544..1F54A; DISALLOWED  # NOTCHED RIGHT SEMICIRCLE WITH THREE DOTS..DO
1F568..1F579; DISALLOWED  # RIGHT SPEAKER..JOYSTICK
1F57B..1F5A3; DISALLOWED  # LEFT HAND TELEPHONE RECEIVER..BLACK DOWN POI
1F5A5..1F5FA; DISALLOWED  # DESKTOP COMPUTER..WORLD MAP
```

```
1F641..1F642; DISALLOWED  # SLIGHTLY FROWNING FACE..SLIGHTLY SMILING FAC
1F650..1F67F; DISALLOWED  # NORTH WEST POINTING LEAF..REVERSE CHECKER BO
1F6C6..1F6CF; DISALLOWED  # TRIANGLE WITH ROUNDED CORNERS..BED
1F6E0..1F6EC; DISALLOWED  # HAMMER AND WRENCH..AIRPLANE ARRIVING
1F6F0..1F6F3; DISALLOWED  # SATELLITE..PASSENGER SHIP
1F780..1F7D4; DISALLOWED  # BLACK LEFT-POINTING ISOSCELES RIGHT TRIANGLE
1F800..1F80B; DISALLOWED  # LEFTWARDS ARROW WITH SMALL TRIANGLE ARROWHEA
1F810..1F847; DISALLOWED  # LEFTWARDS ARROW WITH SMALL EQUILATERAL ARROW
1F850..1F859; DISALLOWED  # LEFTWARDS SANS-SERIF ARROW..UP DOWN SANS-SER
1F860..1F887; DISALLOWED  # WIDE-HEADED LEFTWARDS LIGHT BARB ARROW..WIDE
```

Appendix B. Changes from Unicode 7.0.0 to Unicode 8.0.0

Changes from derived property value UNASSIGNED to either PVALID or DISALLOWED.

```
08B3..08B4  ; PVALID      # ARABIC LETTER AIN WITH THREE DOTS BELOW..ARA
08E3        ; PVALID      # ARABIC TURNED DAMMA BELOW
0AF9        ; PVALID      # GUJARATI LETTER ZHA
0C5A        ; PVALID      # TELUGU LETTER RRRA
0D5F        ; PVALID      # MALAYALAM LETTER ARCHAIC II
13F5        ; PVALID      # CHEROKEE LETTER MV
13F8..13FD  ; DISALLOWED  # CHEROKEE SMALL LETTER YE..CHEROKEE SMALL LET
20BE        ; DISALLOWED  # LARI SIGN
218A..218B  ; DISALLOWED  # TURNED DIGIT TWO..TURNED DIGIT THREE
2BEC..2BEF  ; DISALLOWED  # LEFTWARDS TWO-HEADED ARROW WITH TRIANGLE ARR
9FCD..9FD5  ; PVALID      # <CJK Ideograph>..<CJK Ideograph>
A69E        ; PVALID      # COMBINING CYRILLIC LETTER EF
A78F        ; PVALID      # LATIN LETTER SINOLOGICAL DOT
A7B2..A7B7  ; DISALLOWED  # LATIN CAPITAL LETTER J WITH CROSSED-TAIL..LA
A8FC..A8FD  ; DISALLOWED  # DEVANAGARI SIGN SIDDHAM..DEVANAGARI JAIN OM
AB60..AB63  ; PVALID      # LATIN SMALL LETTER SAKHA YAT..LATIN SMALL LE
AB70..ABBF  ; DISALLOWED  # CHEROKEE SMALL LETTER A..CHEROKEE SMALL LETT
FE2E..FE2F  ; PVALID      # COMBINING CYRILLIC TITLO LEFT HALF..COMBININ
108E0..108F2; PVALID      # HATRAN LETTER ALEPH..HATRAN LETTER QOPH
108F4..108F5; PVALID      # HATRAN LETTER SHIN..HATRAN LETTER TAW
108FB..108FF; DISALLOWED  # HATRAN NUMBER ONE..HATRAN NUMBER ONE HUNDRED
109BC..109BD; DISALLOWED  # MEROITIC CURSIVE FRACTION ELEVEN TWELFTHS..M
109C0..109CF; DISALLOWED  # MEROITIC CURSIVE NUMBER ONE..MEROITIC CURSIV
109D2..109FF; DISALLOWED  # MEROITIC CURSIVE NUMBER ONE HUNDRED..MEROITI
10C80..10CB2; DISALLOWED  # OLD HUNGARIAN CAPITAL LETTER A..OLD HUNGARIA
10CC0..10CF2; PVALID      # OLD HUNGARIAN SMALL LETTER A..OLD HUNGARIAN
10CFA..10CFF; DISALLOWED  # OLD HUNGARIAN NUMBER ONE..OLD HUNGARIAN NUMB
111C9..111CC; DISALLOWED  # SHARADA SANDHI MARK..SHARADA EXTRA SHORT VOW
111DB..111DF; DISALLOWED  # SHARADA SIGN SIDDHAM..SHARADA SECTION MARK-2
11280..11286; PVALID      # MULTANI LETTER A..MULTANI LETTER GA
11288       ; PVALID      # MULTANI LETTER GHA
1128A..1128D; PVALID      # MULTANI LETTER CA..MULTANI LETTER JJA
```

```
1128F..1129D; PVALID      # MULTANI LETTER NYA..MULTANI LETTER BA
1129F..112A9; PVALID      # MULTANI LETTER BHA..MULTANI SECTION MARK
11300       ; PVALID      # GRANTHA SIGN COMBINING ANUSVARA ABOVE
11350       ; PVALID      # GRANTHA OM
115CA..115DD; DISALLOWED  # SIDDHAM SECTION MARK WITH TRIDENT AND U-SHAP
11700..11719; PVALID      # AHOM LETTER KA..AHOM LETTER JHA
1171D..1172B; PVALID      # AHOM CONSONANT SIGN MEDIAL LA..AHOM SIGN KIL
11730..1173F; PVALID      # AHOM DIGIT ZERO..AHOM SYMBOL VI
12399       ; PVALID      # CUNEIFORM SIGN U U
12480..12543; PVALID      # CUNEIFORM SIGN AB TIMES NUN TENU..CUNEIFORM
14400..14646; PVALID      # ANATOLIAN HIEROGLYPH A001..ANATOLIAN HIEROGL
1D1DE..1D1E8; DISALLOWED  # MUSICAL SYMBOL KIEVAN C CLEF..MUSICAL SYMBOL
1D800..1DA8B; DISALLOWED  # SIGNWRITING HAND-FIST INDEX..SIGNWRITING PAR
1DA9B..1DA9F; PVALID      # SIGNWRITING FILL MODIFIER-2..SIGNWRITING FIL
1DAA1..1DAAF; PVALID      # SIGNWRITING ROTATION MODIFIER-2..SIGNWRITING
1F32D..1F32F; DISALLOWED  # HOT DOG..BURRITO
1F37E..1F37F; DISALLOWED  # BOTTLE WITH POPPING CORK..POPCORN
1F3CF..1F3D3; DISALLOWED  # CRICKET BAT AND BALL..TABLE TENNIS PADDLE AN
1F3F8..1F3FF; DISALLOWED  # BADMINTON RACQUET AND SHUTTLECOCK..EMOJI MOD
1F4FF       ; DISALLOWED  # PRAYER BEADS
1F54B..1F54F; DISALLOWED  # KAABA..BOWL OF HYGIEIA
1F643..1F644; DISALLOWED  # UPSIDE-DOWN FACE..FACE WITH ROLLING EYES
1F6D0       ; DISALLOWED  # PLACE OF WORSHIP
1F910..1F918; DISALLOWED  # ZIPPER-MOUTH FACE..SIGN OF THE HORNS
1F980..1F984; DISALLOWED  # CRAB..UNICORN FACE
1F9C0       ; DISALLOWED  # CHEESE WEDGE
```

Appendix C. Changes from Unicode 8.0.0 to Unicode 9.0.0

Changes from derived property value UNASSIGNED to either PVALID or DISALLOWED.

```
08B6..08BD  ; PVALID      # ARABIC LETTER BEH WITH SMALL MEEM ABOVE..ARA
08D4..08E2  ; PVALID      # ARABIC SMALL HIGH WORD AR-RUB..ARABIC DISPUT
0C80        ; PVALID      # KANNADA SIGN SPACING CANDRABINDU
0D4F        ; DISALLOWED  # MALAYALAM SIGN PARA
0D54..0D56  ; PVALID      # MALAYALAM LETTER CHILLU M..MALAYALAM LETTER
0D58..0D5E  ; DISALLOWED  # MALAYALAM FRACTION ONE ONE-HUNDRED-AND-SIXTI
0D76..0D78  ; DISALLOWED  # MALAYALAM FRACTION ONE SIXTEENTH..MALAYALAM
1C80..1C88  ; DISALLOWED  # CYRILLIC SMALL LETTER ROUNDED VE..CYRILLIC S
1DFB        ; PVALID      # COMBINING DELETION MARK
23FB..23FE  ; DISALLOWED  # POWER SYMBOL..POWER SLEEP SYMBOL
2E43..2E44  ; DISALLOWED  # DASH WITH LEFT UPTURN..DOUBLE SUSPENSION MAR
A7AE        ; DISALLOWED  # LATIN CAPITAL LETTER SMALL CAPITAL I
A8C5        ; PVALID      # SAURASHTRA SIGN CANDRABINDU
1018D..1018E; DISALLOWED  # GREEK INDICTION SIGN..NOMISMA SIGN
104B0..104D3; DISALLOWED  # OSAGE CAPITAL LETTER A..OSAGE CAPITAL LETTER
104D8..104FB; PVALID      # OSAGE SMALL LETTER A..OSAGE SMALL LETTER ZHA
```

```
1123E       ; PVALID      # KHOJKI SIGN SUKUN
11400..11459; PVALID      # NEWA LETTER A..NEWA DIGIT NINE
1145B       ; DISALLOWED  # NEWA PLACEHOLDER MARK
1145D       ; DISALLOWED  # NEWA INSERTION SIGN
11660..1166C; DISALLOWED  # MONGOLIAN BIRGA WITH ORNAMENT..MONGOLIAN TUR
11C00..11C08; PVALID      # BHAIKSUKI LETTER A..BHAIKSUKI LETTER VOCALIC
11C0A..11C36; PVALID      # BHAIKSUKI LETTER E..BHAIKSUKI VOWEL SIGN VOC
11C38..11C45; PVALID      # BHAIKSUKI VOWEL SIGN E..BHAIKSUKI GAP FILLER
11C50..11C6C; PVALID      # BHAIKSUKI DIGIT ZERO..BHAIKSUKI HUNDREDS UNI
11C70..11C8F; DISALLOWED  # MARCHEN HEAD MARK..MARCHEN LETTER A
11C92..11CA7; PVALID      # MARCHEN SUBJOINED LETTER KA..MARCHEN SUBJOIN
11CA9..11CB6; PVALID      # MARCHEN SUBJOINED LETTER YA..MARCHEN SIGN CA
16FE0       ; PVALID      # TANGUT ITERATION MARK
17000..187EC; PVALID      # <Tangut Ideograph>..<Tangut Ideograph>
18800..18AF2; PVALID      # TANGUT COMPONENT-001..TANGUT COMPONENT-755
1E000..1E006; PVALID      # COMBINING GLAGOLITIC LETTER AZU..COMBINING G
1E008..1E018; PVALID      # COMBINING GLAGOLITIC LETTER ZEMLJA..COMBININ
1E01B..1E021; PVALID      # COMBINING GLAGOLITIC LETTER SHTA..COMBINING
1E023..1E024; PVALID      # COMBINING GLAGOLITIC LETTER YU..COMBINING GL
1E026..1E02A; PVALID      # COMBINING GLAGOLITIC LETTER YO..COMBINING GL
1E900..1E94A; DISALLOWED  # ADLAM CAPITAL LETTER ALIF..ADLAM NUKTA
1E950..1E959; PVALID      # ADLAM DIGIT ZERO..ADLAM DIGIT NINE
1E95E..1E95F; DISALLOWED  # ADLAM INITIAL EXCLAMATION MARK..ADLAM INITIA
1F19B..1F1AC; DISALLOWED  # SQUARED THREE D..SQUARED VOD
1F23B       ; DISALLOWED  # SQUARED CJK UNIFIED IDEOGRAPH-914D
1F57A       ; DISALLOWED  # MAN DANCING
1F5A4       ; DISALLOWED  # BLACK HEART
1F6D1..1F6D2; DISALLOWED  # OCTAGONAL SIGN..SHOPPING TROLLEY
1F6F4..1F6F6; DISALLOWED  # SCOOTER..CANOE
1F919..1F91E; DISALLOWED  # CALL ME HAND..HAND WITH INDEX AND MIDDLE FIN
1F920..1F927; DISALLOWED  # FACE WITH COWBOY HAT..SNEEZING FACE
1F930       ; DISALLOWED  # PREGNANT WOMAN
1F933..1F93E; DISALLOWED  # SELFIE..HANDBALL
1F940..1F94B; DISALLOWED  # WILTED FLOWER..MARTIAL ARTS UNIFORM
1F950..1F95E; DISALLOWED  # CROISSANT..PANCAKES
```

Appendix D. Changes from Unicode 9.0.0 to Unicode 10.0.0

Changes from derived property value UNASSIGNED to either PVALID or DISALLOWED.

```
0860..086A  ; PVALID      # SYRIAC LETTER MALAYALAM NGA..SYRIAC LETTER M
09FC..09FD  ; PVALID      # BENGALI LETTER VEDIC ANUSVARA..BENGALI ABBRE
0AFA..0AFF  ; PVALID      # GUJARATI SIGN SUKUN..GUJARATI SIGN TWO-CIRCL
0D00        ; PVALID      # MALAYALAM SIGN COMBINING ANUSVARA ABOVE
0D3B..0D3C  ; PVALID      # MALAYALAM SIGN VERTICAL BAR VIRAMA..MALAYALA
1CF7        ; PVALID      # VEDIC SIGN ATIKRAMA
1DF6..1DF9  ; PVALID      # COMBINING KAVYKA ABOVE RIGHT..COMBINING WIDE
20BF        ; DISALLOWED  # BITCOIN SIGN
23FF        ; DISALLOWED  # OBSERVER EYE SYMBOL
2BD2        ; DISALLOWED  # GROUP MARK
2E45..2E49  ; DISALLOWED  # INVERTED LOW KAVYKA..DOUBLE STACKED COMMA
312E        ; PVALID      # BOPOMOFO LETTER O WITH DOT ABOVE
9FD6..9FEA  ; PVALID      # <CJK Ideograph>..<CJK Ideograph>
1032D..1032F; PVALID      # OLD ITALIC LETTER YE..OLD ITALIC LETTER SOUT
11A00..11A47; PVALID      # ZANABAZAR SQUARE LETTER A..ZANABAZAR SQUARE
11A50..11A83; PVALID      # SOYOMBO LETTER A..SOYOMBO LETTER KSSA
11A86..11A9C; PVALID      # SOYOMBO CLUSTER-INITIAL LETTER RA..SOYOMBO M
11A9E..11AA2; DISALLOWED  # SOYOMBO HEAD MARK WITH MOON AND SUN AND TRIP
11D00..11D06; PVALID      # MASARAM GONDI LETTER A..MASARAM GONDI LETTER
11D08..11D09; PVALID      # MASARAM GONDI LETTER AI..MASARAM GONDI LETTE
11D0B..11D36; PVALID      # MASARAM GONDI LETTER AU..MASARAM GONDI VOWEL
11D3A       ; PVALID      # MASARAM GONDI VOWEL SIGN E
11D3C..11D3D; PVALID      # MASARAM GONDI VOWEL SIGN AI..MASARAM GONDI V
11D3F..11D47; PVALID      # MASARAM GONDI VOWEL SIGN AU..MASARAM GONDI R
11D50..11D59; PVALID      # MASARAM GONDI DIGIT ZERO..MASARAM GONDI DIGI
16FE1       ; PVALID      # NUSHU ITERATION MARK
1B002..1B11E; PVALID      # HENTAIGANA LETTER A-1..HENTAIGANA LETTER N-M
1B170..1B2FB; PVALID      # NUSHU CHARACTER-1B170..NUSHU CHARACTER-1B2FB
1F260..1F265; DISALLOWED  # ROUNDED SYMBOL FOR FU..ROUNDED SYMBOL FOR CA
1F6D3..1F6D4; DISALLOWED  # STUPA..PAGODA
1F6F7..1F6F8; DISALLOWED  # SLED..FLYING SAUCER
1F900..1F90B; DISALLOWED  # CIRCLED CROSS FORMEE WITH FOUR DOTS..DOWNWAR
1F91F       ; DISALLOWED  # I LOVE YOU HAND SIGN
1F928..1F92F; DISALLOWED  # FACE WITH ONE EYEBROW RAISED..SHOCKED FACE W
1F931..1F932; DISALLOWED  # BREAST-FEEDING..PALMS UP TOGETHER
1F94C       ; DISALLOWED  # CURLING STONE
1F95F..1F96B; DISALLOWED  # DUMPLING..CANNED FOOD
1F992..1F997; DISALLOWED  # GIRAFFE FACE..CRICKET
1F9D0..1F9E6; DISALLOWED  # FACE WITH MONOCLE..SOCKS
```

Appendix E. Changes from Unicode 10.0.0 to Unicode 11.0.0

Changes from derived property value DISALLOWED to PVALID.

```
111C9       ; PVALID      # SHARADA SANDHI MARK
```

Changes from derived property value UNASSIGNED to either PVALID or DISALLOWED.

```
0560          ; PVALID      # ARMENIAN SMALL LETTER TURNED AYB
0588          ; PVALID      # ARMENIAN SMALL LETTER YI WITH STROKE
05EF          ; PVALID      # HEBREW YOD TRIANGLE
07FD..07FF    ; PVALID      # NKO DANTAYALAN..NKO TAMAN SIGN
08D3          ; PVALID      # ARABIC SMALL LOW WAW
09FE          ; PVALID      # BENGALI SANDHI MARK
0A76          ; DISALLOWED  # GURMUKHI ABBREVIATION SIGN
0C04          ; PVALID      # TELUGU SIGN COMBINING ANUSVARA ABOVE
0C84          ; DISALLOWED  # KANNADA SIGN SIDDHAM
1878          ; PVALID      # MONGOLIAN LETTER CHA WITH TWO DOTS
1C90..1CBA    ; DISALLOWED  # GEORGIAN MTAVRULI CAPITAL LETTER AN..GEORGIA
1CBD..1CBF    ; DISALLOWED  # GEORGIAN MTAVRULI CAPITAL LETTER AEN..GEORGI
2BBA..2BBC    ; DISALLOWED  # OVERLAPPING WHITE SQUARES..OVERLAPPING BLACK
2BD3..2BEB    ; DISALLOWED  # PLUTO FORM TWO..STAR WITH RIGHT HALF BLACK
2BF0..2BFE    ; DISALLOWED  # ERIS FORM ONE..REVERSED RIGHT ANGLE
2E4A..2E4E    ; DISALLOWED  # DOTTED SOLIDUS..PUNCTUS ELEVATUS MARK
312F          ; PVALID      # BOPOMOFO LETTER NN
9FEB..9FEF    ; PVALID      # <CJK Ideograph>..<CJK Ideograph>
A7AF          ; PVALID      # LATIN LETTER SMALL CAPITAL Q
A7B8..A7B9    ; DISALLOWED  # LATIN CAPITAL LETTER U WITH STROKE..LATIN SM
A8FE..A8FF    ; PVALID      # DEVANAGARI LETTER AY..DEVANAGARI VOWEL SIGN
10A34..10A35; PVALID        # KHAROSHTHI LETTER TTTA..KHAROSHTHI LETTER VH
10A48         ; DISALLOWED  # KHAROSHTHI FRACTION ONE HALF
10D00..10D27; PVALID        # HANIFI ROHINGYA LETTER A..HANIFI ROHINGYA SI
10D30..10D39; PVALID        # HANIFI ROHINGYA DIGIT ZERO..HANIFI ROHINGYA
10F00..10F27; PVALID        # OLD SOGDIAN LETTER ALEPH..OLD SOGDIAN LIGATU
10F30..10F59; PVALID        # SOGDIAN LETTER ALEPH..SOGDIAN PUNCTUATION HA
110CD         ; DISALLOWED  # KAITHI NUMBER SIGN ABOVE
11144..11146; PVALID        # CHAKMA LETTER LHAA..CHAKMA VOWEL SIGN EI
1133B         ; PVALID      # COMBINING BINDU BELOW
1145E         ; PVALID      # NEWA SANDHI MARK
1171A         ; PVALID      # AHOM LETTER ALTERNATE BA
11800..1183B; PVALID        # DOGRA LETTER A..DOGRA ABBREVIATION SIGN
11A9D         ; PVALID      # SOYOMBO MARK PLUTA
11D60..11D65; PVALID        # GUNJALA GONDI LETTER A..GUNJALA GONDI LETTER
11D67..11D68; PVALID        # GUNJALA GONDI LETTER EE..GUNJALA GONDI LETTE
11D6A..11D8E; PVALID        # GUNJALA GONDI LETTER OO..GUNJALA GONDI VOWEL
11D90..11D91; PVALID        # GUNJALA GONDI VOWEL SIGN EE..GUNJALA GONDI V
11D93..11D98; PVALID        # GUNJALA GONDI VOWEL SIGN OO..GUNJALA GONDI O
11DA0..11DA9; PVALID        # GUNJALA GONDI DIGIT ZERO..GUNJALA GONDI DIGI
11EE0..11EF8; PVALID        # MAKASAR LETTER KA..MAKASAR END OF SECTION
16E40..16E9A; DISALLOWED    # MEDEFAIDRIN CAPITAL LETTER M..MEDEFAIDRIN EX
187ED..187F1; PVALID        # <Tangut Ideograph>..<Tangut Ideograph>
1D2E0..1D2F3; DISALLOWED    # MAYAN NUMERAL ZERO..MAYAN NUMERAL NINETEEN
1D372..1D378; DISALLOWED    # IDEOGRAPHIC TALLY MARK ONE..TALLY MARK FIVE
1EC71..1ECB4; DISALLOWED    # INDIC SIYAQ NUMBER ONE..INDIC SIYAQ ALTERNAT
1F12F         ; DISALLOWED  # COPYLEFT SYMBOL
1F6F9         ; DISALLOWED  # SKATEBOARD
```

```
1F7D5..1F7D8; DISALLOWED  # CIRCLED TRIANGLE..NEGATIVE CIRCLED SQUARE
1F94D..1F94F; DISALLOWED  # LACROSSE STICK AND BALL..FLYING DISC
1F96C..1F970; DISALLOWED  # LEAFY GREEN..SMILING FACE WITH SMILING EYES
1F973..1F976; DISALLOWED  # FACE WITH PARTY HORN AND PARTY HAT..FREEZING
1F97A       ; DISALLOWED  # FACE WITH PLEADING EYES
1F97C..1F97F; DISALLOWED  # LAB COAT..FLAT SHOE
1F998..1F9A2; DISALLOWED  # KANGAROO..SWAN
1F9B0..1F9B9; DISALLOWED  # EMOJI COMPONENT RED HAIR..SUPERVILLAIN
1F9C1..1F9C2; DISALLOWED  # CUPCAKE..SALT SHAKER
1F9E7..1F9FF; DISALLOWED  # RED GIFT ENVELOPE..NAZAR AMULET
```

Appendix F. Code points in Unicode Character Database (UCD) format for Unicode 11.0.0

```
0000..002C  ; DISALLOWED  # <control>..COMMA
002D        ; PVALID      # HYPHEN-MINUS
002E..002F  ; DISALLOWED  # FULL STOP..SOLIDUS
0030..0039  ; PVALID      # DIGIT ZERO..DIGIT NINE
003A..0060  ; DISALLOWED  # COLON..GRAVE ACCENT
0061..007A  ; PVALID      # LATIN SMALL LETTER A..LATIN SMALL LETTER Z
007B..00B6  ; DISALLOWED  # LEFT CURLY BRACKET..PILCROW SIGN
00B7        ; CONTEXTO    # MIDDLE DOT
00B8..00DE  ; DISALLOWED  # CEDILLA..LATIN CAPITAL LETTER THORN
00DF..00F6  ; PVALID      # LATIN SMALL LETTER SHARP S..LATIN SMALL LETT
00F7        ; DISALLOWED  # DIVISION SIGN
00F8..00FF  ; PVALID      # LATIN SMALL LETTER O WITH STROKE..LATIN SMAL
0100        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH MACRON
0101        ; PVALID      # LATIN SMALL LETTER A WITH MACRON
0102        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH BREVE
0103        ; PVALID      # LATIN SMALL LETTER A WITH BREVE
0104        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH OGONEK
0105        ; PVALID      # LATIN SMALL LETTER A WITH OGONEK
0106        ; DISALLOWED  # LATIN CAPITAL LETTER C WITH ACUTE
0107        ; PVALID      # LATIN SMALL LETTER C WITH ACUTE
0108        ; DISALLOWED  # LATIN CAPITAL LETTER C WITH CIRCUMFLEX
0109        ; PVALID      # LATIN SMALL LETTER C WITH CIRCUMFLEX
010A        ; DISALLOWED  # LATIN CAPITAL LETTER C WITH DOT ABOVE
010B        ; PVALID      # LATIN SMALL LETTER C WITH DOT ABOVE
010C        ; DISALLOWED  # LATIN CAPITAL LETTER C WITH CARON
010D        ; PVALID      # LATIN SMALL LETTER C WITH CARON
010E        ; DISALLOWED  # LATIN CAPITAL LETTER D WITH CARON
010F        ; PVALID      # LATIN SMALL LETTER D WITH CARON
0110        ; DISALLOWED  # LATIN CAPITAL LETTER D WITH STROKE
0111        ; PVALID      # LATIN SMALL LETTER D WITH STROKE
0112        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH MACRON
0113        ; PVALID      # LATIN SMALL LETTER E WITH MACRON
0114        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH BREVE
0115        ; PVALID      # LATIN SMALL LETTER E WITH BREVE
```

```
0116        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH DOT ABOVE
0117        ; PVALID      # LATIN SMALL LETTER E WITH DOT ABOVE
0118        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH OGONEK
0119        ; PVALID      # LATIN SMALL LETTER E WITH OGONEK
011A        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH CARON
011B        ; PVALID      # LATIN SMALL LETTER E WITH CARON
011C        ; DISALLOWED  # LATIN CAPITAL LETTER G WITH CIRCUMFLEX
011D        ; PVALID      # LATIN SMALL LETTER G WITH CIRCUMFLEX
011E        ; DISALLOWED  # LATIN CAPITAL LETTER G WITH BREVE
011F        ; PVALID      # LATIN SMALL LETTER G WITH BREVE
0120        ; DISALLOWED  # LATIN CAPITAL LETTER G WITH DOT ABOVE
0121        ; PVALID      # LATIN SMALL LETTER G WITH DOT ABOVE
0122        ; DISALLOWED  # LATIN CAPITAL LETTER G WITH CEDILLA
0123        ; PVALID      # LATIN SMALL LETTER G WITH CEDILLA
0124        ; DISALLOWED  # LATIN CAPITAL LETTER H WITH CIRCUMFLEX
0125        ; PVALID      # LATIN SMALL LETTER H WITH CIRCUMFLEX
0126        ; DISALLOWED  # LATIN CAPITAL LETTER H WITH STROKE
0127        ; PVALID      # LATIN SMALL LETTER H WITH STROKE
0128        ; DISALLOWED  # LATIN CAPITAL LETTER I WITH TILDE
0129        ; PVALID      # LATIN SMALL LETTER I WITH TILDE
012A        ; DISALLOWED  # LATIN CAPITAL LETTER I WITH MACRON
012B        ; PVALID      # LATIN SMALL LETTER I WITH MACRON
012C        ; DISALLOWED  # LATIN CAPITAL LETTER I WITH BREVE
012D        ; PVALID      # LATIN SMALL LETTER I WITH BREVE
012E        ; DISALLOWED  # LATIN CAPITAL LETTER I WITH OGONEK
012F        ; PVALID      # LATIN SMALL LETTER I WITH OGONEK
0130        ; DISALLOWED  # LATIN CAPITAL LETTER I WITH DOT ABOVE
0131        ; PVALID      # LATIN SMALL LETTER DOTLESS I
0132..0134  ; DISALLOWED  # LATIN CAPITAL LIGATURE IJ..LATIN CAPITAL LET
0135        ; PVALID      # LATIN SMALL LETTER J WITH CIRCUMFLEX
0136        ; DISALLOWED  # LATIN CAPITAL LETTER K WITH CEDILLA
0137..0138  ; PVALID      # LATIN SMALL LETTER K WITH CEDILLA..LATIN SMA
0139        ; DISALLOWED  # LATIN CAPITAL LETTER L WITH ACUTE
013A        ; PVALID      # LATIN SMALL LETTER L WITH ACUTE
013B        ; DISALLOWED  # LATIN CAPITAL LETTER L WITH CEDILLA
013C        ; PVALID      # LATIN SMALL LETTER L WITH CEDILLA
013D        ; DISALLOWED  # LATIN CAPITAL LETTER L WITH CARON
013E        ; PVALID      # LATIN SMALL LETTER L WITH CARON
013F..0141  ; DISALLOWED  # LATIN CAPITAL LETTER L WITH MIDDLE DOT..LATI
0142        ; PVALID      # LATIN SMALL LETTER L WITH STROKE
0143        ; DISALLOWED  # LATIN CAPITAL LETTER N WITH ACUTE
0144        ; PVALID      # LATIN SMALL LETTER N WITH ACUTE
0145        ; DISALLOWED  # LATIN CAPITAL LETTER N WITH CEDILLA
0146        ; PVALID      # LATIN SMALL LETTER N WITH CEDILLA
0147        ; DISALLOWED  # LATIN CAPITAL LETTER N WITH CARON
0148        ; PVALID      # LATIN SMALL LETTER N WITH CARON
0149..014A  ; DISALLOWED  # LATIN SMALL LETTER N PRECEDED BY APOSTROPHE.
014B        ; PVALID      # LATIN SMALL LETTER ENG
```

```
014C        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH MACRON
014D        ; PVALID      # LATIN SMALL LETTER O WITH MACRON
014E        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH BREVE
014F        ; PVALID      # LATIN SMALL LETTER O WITH BREVE
0150        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH DOUBLE ACUTE
0151        ; PVALID      # LATIN SMALL LETTER O WITH DOUBLE ACUTE
0152        ; DISALLOWED  # LATIN CAPITAL LIGATURE OE
0153        ; PVALID      # LATIN SMALL LIGATURE OE
0154        ; DISALLOWED  # LATIN CAPITAL LETTER R WITH ACUTE
0155        ; PVALID      # LATIN SMALL LETTER R WITH ACUTE
0156        ; DISALLOWED  # LATIN CAPITAL LETTER R WITH CEDILLA
0157        ; PVALID      # LATIN SMALL LETTER R WITH CEDILLA
0158        ; DISALLOWED  # LATIN CAPITAL LETTER R WITH CARON
0159        ; PVALID      # LATIN SMALL LETTER R WITH CARON
015A        ; DISALLOWED  # LATIN CAPITAL LETTER S WITH ACUTE
015B        ; PVALID      # LATIN SMALL LETTER S WITH ACUTE
015C        ; DISALLOWED  # LATIN CAPITAL LETTER S WITH CIRCUMFLEX
015D        ; PVALID      # LATIN SMALL LETTER S WITH CIRCUMFLEX
015E        ; DISALLOWED  # LATIN CAPITAL LETTER S WITH CEDILLA
015F        ; PVALID      # LATIN SMALL LETTER S WITH CEDILLA
0160        ; DISALLOWED  # LATIN CAPITAL LETTER S WITH CARON
0161        ; PVALID      # LATIN SMALL LETTER S WITH CARON
0162        ; DISALLOWED  # LATIN CAPITAL LETTER T WITH CEDILLA
0163        ; PVALID      # LATIN SMALL LETTER T WITH CEDILLA
0164        ; DISALLOWED  # LATIN CAPITAL LETTER T WITH CARON
0165        ; PVALID      # LATIN SMALL LETTER T WITH CARON
0166        ; DISALLOWED  # LATIN CAPITAL LETTER T WITH STROKE
0167        ; PVALID      # LATIN SMALL LETTER T WITH STROKE
0168        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH TILDE
0169        ; PVALID      # LATIN SMALL LETTER U WITH TILDE
016A        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH MACRON
016B        ; PVALID      # LATIN SMALL LETTER U WITH MACRON
016C        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH BREVE
016D        ; PVALID      # LATIN SMALL LETTER U WITH BREVE
016E        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH RING ABOVE
016F        ; PVALID      # LATIN SMALL LETTER U WITH RING ABOVE
0170        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH DOUBLE ACUTE
0171        ; PVALID      # LATIN SMALL LETTER U WITH DOUBLE ACUTE
0172        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH OGONEK
0173        ; PVALID      # LATIN SMALL LETTER U WITH OGONEK
0174        ; DISALLOWED  # LATIN CAPITAL LETTER W WITH CIRCUMFLEX
0175        ; PVALID      # LATIN SMALL LETTER W WITH CIRCUMFLEX
0176        ; DISALLOWED  # LATIN CAPITAL LETTER Y WITH CIRCUMFLEX
0177        ; PVALID      # LATIN SMALL LETTER Y WITH CIRCUMFLEX
0178..0179  ; DISALLOWED  # LATIN CAPITAL LETTER Y WITH DIAERESIS..LATIN
017A        ; PVALID      # LATIN SMALL LETTER Z WITH ACUTE
017B        ; DISALLOWED  # LATIN CAPITAL LETTER Z WITH DOT ABOVE
017C        ; PVALID      # LATIN SMALL LETTER Z WITH DOT ABOVE
```

```
017D        ; DISALLOWED  # LATIN CAPITAL LETTER Z WITH CARON
017E        ; PVALID      # LATIN SMALL LETTER Z WITH CARON
017F        ; DISALLOWED  # LATIN SMALL LETTER LONG S
0180        ; PVALID      # LATIN SMALL LETTER B WITH STROKE
0181..0182  ; DISALLOWED  # LATIN CAPITAL LETTER B WITH HOOK..LATIN CAPI
0183        ; PVALID      # LATIN SMALL LETTER B WITH TOPBAR
0184        ; DISALLOWED  # LATIN CAPITAL LETTER TONE SIX
0185        ; PVALID      # LATIN SMALL LETTER TONE SIX
0186..0187  ; DISALLOWED  # LATIN CAPITAL LETTER OPEN O..LATIN CAPITAL L
0188        ; PVALID      # LATIN SMALL LETTER C WITH HOOK
0189..018B  ; DISALLOWED  # LATIN CAPITAL LETTER AFRICAN D..LATIN CAPITA
018C..018D  ; PVALID      # LATIN SMALL LETTER D WITH TOPBAR..LATIN SMAL
018E..0191  ; DISALLOWED  # LATIN CAPITAL LETTER REVERSED E..LATIN CAPIT
0192        ; PVALID      # LATIN SMALL LETTER F WITH HOOK
0193..0194  ; DISALLOWED  # LATIN CAPITAL LETTER G WITH HOOK..LATIN CAPI
0195        ; PVALID      # LATIN SMALL LETTER HV
0196..0198  ; DISALLOWED  # LATIN CAPITAL LETTER IOTA..LATIN CAPITAL LET
0199..019B  ; PVALID      # LATIN SMALL LETTER K WITH HOOK..LATIN SMALL
019C..019D  ; DISALLOWED  # LATIN CAPITAL LETTER TURNED M..LATIN CAPITAL
019E        ; PVALID      # LATIN SMALL LETTER N WITH LONG RIGHT LEG
019F..01A0  ; DISALLOWED  # LATIN CAPITAL LETTER O WITH MIDDLE TILDE..LA
01A1        ; PVALID      # LATIN SMALL LETTER O WITH HORN
01A2        ; DISALLOWED  # LATIN CAPITAL LETTER OI
01A3        ; PVALID      # LATIN SMALL LETTER OI
01A4        ; DISALLOWED  # LATIN CAPITAL LETTER P WITH HOOK
01A5        ; PVALID      # LATIN SMALL LETTER P WITH HOOK
01A6..01A7  ; DISALLOWED  # LATIN LETTER YR..LATIN CAPITAL LETTER TONE T
01A8        ; PVALID      # LATIN SMALL LETTER TONE TWO
01A9        ; DISALLOWED  # LATIN CAPITAL LETTER ESH
01AA..01AB  ; PVALID      # LATIN LETTER REVERSED ESH LOOP..LATIN SMALL
01AC        ; DISALLOWED  # LATIN CAPITAL LETTER T WITH HOOK
01AD        ; PVALID      # LATIN SMALL LETTER T WITH HOOK
01AE..01AF  ; DISALLOWED  # LATIN CAPITAL LETTER T WITH RETROFLEX HOOK..
01B0        ; PVALID      # LATIN SMALL LETTER U WITH HORN
01B1..01B3  ; DISALLOWED  # LATIN CAPITAL LETTER UPSILON..LATIN CAPITAL
01B4        ; PVALID      # LATIN SMALL LETTER Y WITH HOOK
01B5        ; DISALLOWED  # LATIN CAPITAL LETTER Z WITH STROKE
01B6        ; PVALID      # LATIN SMALL LETTER Z WITH STROKE
01B7..01B8  ; DISALLOWED  # LATIN CAPITAL LETTER EZH..LATIN CAPITAL LETT
01B9..01BB  ; PVALID      # LATIN SMALL LETTER EZH REVERSED..LATIN LETTE
01BC        ; DISALLOWED  # LATIN CAPITAL LETTER TONE FIVE
01BD..01C3  ; PVALID      # LATIN SMALL LETTER TONE FIVE..LATIN LETTER R
01C4..01CD  ; DISALLOWED  # LATIN CAPITAL LETTER DZ WITH CARON..LATIN CA
01CE        ; PVALID      # LATIN SMALL LETTER A WITH CARON
01CF        ; DISALLOWED  # LATIN CAPITAL LETTER I WITH CARON
01D0        ; PVALID      # LATIN SMALL LETTER I WITH CARON
01D1        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH CARON
01D2        ; PVALID      # LATIN SMALL LETTER O WITH CARON
```

```
01D3        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH CARON
01D4        ; PVALID      # LATIN SMALL LETTER U WITH CARON
01D5        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH DIAERESIS AND MA
01D6        ; PVALID      # LATIN SMALL LETTER U WITH DIAERESIS AND MACR
01D7        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH DIAERESIS AND AC
01D8        ; PVALID      # LATIN SMALL LETTER U WITH DIAERESIS AND ACUT
01D9        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH DIAERESIS AND CA
01DA        ; PVALID      # LATIN SMALL LETTER U WITH DIAERESIS AND CARO
01DB        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH DIAERESIS AND GR
01DC..01DD  ; PVALID      # LATIN SMALL LETTER U WITH DIAERESIS AND GRAV
01DE        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH DIAERESIS AND MA
01DF        ; PVALID      # LATIN SMALL LETTER A WITH DIAERESIS AND MACR
01E0        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH DOT ABOVE AND MA
01E1        ; PVALID      # LATIN SMALL LETTER A WITH DOT ABOVE AND MACR
01E2        ; DISALLOWED  # LATIN CAPITAL LETTER AE WITH MACRON
01E3        ; PVALID      # LATIN SMALL LETTER AE WITH MACRON
01E4        ; DISALLOWED  # LATIN CAPITAL LETTER G WITH STROKE
01E5        ; PVALID      # LATIN SMALL LETTER G WITH STROKE
01E6        ; DISALLOWED  # LATIN CAPITAL LETTER G WITH CARON
01E7        ; PVALID      # LATIN SMALL LETTER G WITH CARON
01E8        ; DISALLOWED  # LATIN CAPITAL LETTER K WITH CARON
01E9        ; PVALID      # LATIN SMALL LETTER K WITH CARON
01EA        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH OGONEK
01EB        ; PVALID      # LATIN SMALL LETTER O WITH OGONEK
01EC        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH OGONEK AND MACRO
01ED        ; PVALID      # LATIN SMALL LETTER O WITH OGONEK AND MACRON
01EE        ; DISALLOWED  # LATIN CAPITAL LETTER EZH WITH CARON
01EF..01F0  ; PVALID      # LATIN SMALL LETTER EZH WITH CARON..LATIN SMA
01F1..01F4  ; DISALLOWED  # LATIN CAPITAL LETTER DZ..LATIN CAPITAL LETTE
01F5        ; PVALID      # LATIN SMALL LETTER G WITH ACUTE
01F6..01F8  ; DISALLOWED  # LATIN CAPITAL LETTER HWAIR..LATIN CAPITAL LE
01F9        ; PVALID      # LATIN SMALL LETTER N WITH GRAVE
01FA        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH RING ABOVE AND A
01FB        ; PVALID      # LATIN SMALL LETTER A WITH RING ABOVE AND ACU
01FC        ; DISALLOWED  # LATIN CAPITAL LETTER AE WITH ACUTE
01FD        ; PVALID      # LATIN SMALL LETTER AE WITH ACUTE
01FE        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH STROKE AND ACUTE
01FF        ; PVALID      # LATIN SMALL LETTER O WITH STROKE AND ACUTE
0200        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH DOUBLE GRAVE
0201        ; PVALID      # LATIN SMALL LETTER A WITH DOUBLE GRAVE
0202        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH INVERTED BREVE
0203        ; PVALID      # LATIN SMALL LETTER A WITH INVERTED BREVE
0204        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH DOUBLE GRAVE
0205        ; PVALID      # LATIN SMALL LETTER E WITH DOUBLE GRAVE
0206        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH INVERTED BREVE
0207        ; PVALID      # LATIN SMALL LETTER E WITH INVERTED BREVE
0208        ; DISALLOWED  # LATIN CAPITAL LETTER I WITH DOUBLE GRAVE
0209        ; PVALID      # LATIN SMALL LETTER I WITH DOUBLE GRAVE
```

```
020A        ; DISALLOWED  # LATIN CAPITAL LETTER I WITH INVERTED BREVE
020B        ; PVALID      # LATIN SMALL LETTER I WITH INVERTED BREVE
020C        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH DOUBLE GRAVE
020D        ; PVALID      # LATIN SMALL LETTER O WITH DOUBLE GRAVE
020E        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH INVERTED BREVE
020F        ; PVALID      # LATIN SMALL LETTER O WITH INVERTED BREVE
0210        ; DISALLOWED  # LATIN CAPITAL LETTER R WITH DOUBLE GRAVE
0211        ; PVALID      # LATIN SMALL LETTER R WITH DOUBLE GRAVE
0212        ; DISALLOWED  # LATIN CAPITAL LETTER R WITH INVERTED BREVE
0213        ; PVALID      # LATIN SMALL LETTER R WITH INVERTED BREVE
0214        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH DOUBLE GRAVE
0215        ; PVALID      # LATIN SMALL LETTER U WITH DOUBLE GRAVE
0216        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH INVERTED BREVE
0217        ; PVALID      # LATIN SMALL LETTER U WITH INVERTED BREVE
0218        ; DISALLOWED  # LATIN CAPITAL LETTER S WITH COMMA BELOW
0219        ; PVALID      # LATIN SMALL LETTER S WITH COMMA BELOW
021A        ; DISALLOWED  # LATIN CAPITAL LETTER T WITH COMMA BELOW
021B        ; PVALID      # LATIN SMALL LETTER T WITH COMMA BELOW
021C        ; DISALLOWED  # LATIN CAPITAL LETTER YOGH
021D        ; PVALID      # LATIN SMALL LETTER YOGH
021E        ; DISALLOWED  # LATIN CAPITAL LETTER H WITH CARON
021F        ; PVALID      # LATIN SMALL LETTER H WITH CARON
0220        ; DISALLOWED  # LATIN CAPITAL LETTER N WITH LONG RIGHT LEG
0221        ; PVALID      # LATIN SMALL LETTER D WITH CURL
0222        ; DISALLOWED  # LATIN CAPITAL LETTER OU
0223        ; PVALID      # LATIN SMALL LETTER OU
0224        ; DISALLOWED  # LATIN CAPITAL LETTER Z WITH HOOK
0225        ; PVALID      # LATIN SMALL LETTER Z WITH HOOK
0226        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH DOT ABOVE
0227        ; PVALID      # LATIN SMALL LETTER A WITH DOT ABOVE
0228        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH CEDILLA
0229        ; PVALID      # LATIN SMALL LETTER E WITH CEDILLA
022A        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH DIAERESIS AND MA
022B        ; PVALID      # LATIN SMALL LETTER O WITH DIAERESIS AND MACR
022C        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH TILDE AND MACRON
022D        ; PVALID      # LATIN SMALL LETTER O WITH TILDE AND MACRON
022E        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH DOT ABOVE
022F        ; PVALID      # LATIN SMALL LETTER O WITH DOT ABOVE
0230        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH DOT ABOVE AND MA
0231        ; PVALID      # LATIN SMALL LETTER O WITH DOT ABOVE AND MACR
0232        ; DISALLOWED  # LATIN CAPITAL LETTER Y WITH MACRON
0233..0239  ; PVALID      # LATIN SMALL LETTER Y WITH MACRON..LATIN SMAL
023A..023B  ; DISALLOWED  # LATIN CAPITAL LETTER A WITH STROKE..LATIN CA
023C        ; PVALID      # LATIN SMALL LETTER C WITH STROKE
023D..023E  ; DISALLOWED  # LATIN CAPITAL LETTER L WITH BAR..LATIN CAPIT
023F..0240  ; PVALID      # LATIN SMALL LETTER S WITH SWASH TAIL..LATIN
0241        ; DISALLOWED  # LATIN CAPITAL LETTER GLOTTAL STOP
0242        ; PVALID      # LATIN SMALL LETTER GLOTTAL STOP
```

```
0243..0246  ; DISALLOWED  # LATIN CAPITAL LETTER B WITH STROKE..LATIN CA
0247        ; PVALID      # LATIN SMALL LETTER E WITH STROKE
0248        ; DISALLOWED  # LATIN CAPITAL LETTER J WITH STROKE
0249        ; PVALID      # LATIN SMALL LETTER J WITH STROKE
024A        ; DISALLOWED  # LATIN CAPITAL LETTER SMALL Q WITH HOOK TAIL
024B        ; PVALID      # LATIN SMALL LETTER Q WITH HOOK TAIL
024C        ; DISALLOWED  # LATIN CAPITAL LETTER R WITH STROKE
024D        ; PVALID      # LATIN SMALL LETTER R WITH STROKE
024E        ; DISALLOWED  # LATIN CAPITAL LETTER Y WITH STROKE
024F..02AF  ; PVALID      # LATIN SMALL LETTER Y WITH STROKE..LATIN SMAL
02B0..02B8  ; DISALLOWED  # MODIFIER LETTER SMALL H..MODIFIER LETTER SMA
02B9..02C1  ; PVALID      # MODIFIER LETTER PRIME..MODIFIER LETTER REVER
02C2..02C5  ; DISALLOWED  # MODIFIER LETTER LEFT ARROWHEAD..MODIFIER LET
02C6..02D1  ; PVALID      # MODIFIER LETTER CIRCUMFLEX ACCENT..MODIFIER
02D2..02EB  ; DISALLOWED  # MODIFIER LETTER CENTRED RIGHT HALF RING..MOD
02EC        ; PVALID      # MODIFIER LETTER VOICING
02ED        ; DISALLOWED  # MODIFIER LETTER UNASPIRATED
02EE        ; PVALID      # MODIFIER LETTER DOUBLE APOSTROPHE
02EF..02FF  ; DISALLOWED  # MODIFIER LETTER LOW DOWN ARROWHEAD..MODIFIER
0300..033F  ; PVALID      # COMBINING GRAVE ACCENT..COMBINING DOUBLE OVE
0340..0341  ; DISALLOWED  # COMBINING GRAVE TONE MARK..COMBINING ACUTE T
0342        ; PVALID      # COMBINING GREEK PERISPOMENI
0343..0345  ; DISALLOWED  # COMBINING GREEK KORONIS..COMBINING GREEK YPO
0346..034E  ; PVALID      # COMBINING BRIDGE ABOVE..COMBINING UPWARDS AR
034F        ; DISALLOWED  # COMBINING GRAPHEME JOINER
0350..036F  ; PVALID      # COMBINING RIGHT ARROWHEAD ABOVE..COMBINING L
0370        ; DISALLOWED  # GREEK CAPITAL LETTER HETA
0371        ; PVALID      # GREEK SMALL LETTER HETA
0372        ; DISALLOWED  # GREEK CAPITAL LETTER ARCHAIC SAMPI
0373        ; PVALID      # GREEK SMALL LETTER ARCHAIC SAMPI
0374        ; DISALLOWED  # GREEK NUMERAL SIGN
0375        ; CONTEXTO    # GREEK LOWER NUMERAL SIGN
0376        ; DISALLOWED  # GREEK CAPITAL LETTER PAMPHYLIAN DIGAMMA
0377        ; PVALID      # GREEK SMALL LETTER PAMPHYLIAN DIGAMMA
0378..0379  ; UNASSIGNED  # <reserved>..<reserved>
037A        ; DISALLOWED  # GREEK YPOGEGRAMMENI
037B..037D  ; PVALID      # GREEK SMALL REVERSED LUNATE SIGMA SYMBOL..GR
037E..037F  ; DISALLOWED  # GREEK QUESTION MARK..GREEK CAPITAL LETTER YO
0380..0383  ; UNASSIGNED  # <reserved>..<reserved>
0384..038A  ; DISALLOWED  # GREEK TONOS..GREEK CAPITAL LETTER IOTA WITH
038B        ; UNASSIGNED  # <reserved>
038C        ; DISALLOWED  # GREEK CAPITAL LETTER OMICRON WITH TONOS
038D        ; UNASSIGNED  # <reserved>
038E..038F  ; DISALLOWED  # GREEK CAPITAL LETTER UPSILON WITH TONOS..GRE
0390        ; PVALID      # GREEK SMALL LETTER IOTA WITH DIALYTIKA AND T
0391..03A1  ; DISALLOWED  # GREEK CAPITAL LETTER ALPHA..GREEK CAPITAL LE
03A2        ; UNASSIGNED  # <reserved>
03A3..03AB  ; DISALLOWED  # GREEK CAPITAL LETTER SIGMA..GREEK CAPITAL LE
```

```
03AC..03CE  ; PVALID      # GREEK SMALL LETTER ALPHA WITH TONOS..GREEK S
03CF..03D6  ; DISALLOWED  # GREEK CAPITAL KAI SYMBOL..GREEK PI SYMBOL
03D7        ; PVALID      # GREEK KAI SYMBOL
03D8        ; DISALLOWED  # GREEK LETTER ARCHAIC KOPPA
03D9        ; PVALID      # GREEK SMALL LETTER ARCHAIC KOPPA
03DA        ; DISALLOWED  # GREEK LETTER STIGMA
03DB        ; PVALID      # GREEK SMALL LETTER STIGMA
03DC        ; DISALLOWED  # GREEK LETTER DIGAMMA
03DD        ; PVALID      # GREEK SMALL LETTER DIGAMMA
03DE        ; DISALLOWED  # GREEK LETTER KOPPA
03DF        ; PVALID      # GREEK SMALL LETTER KOPPA
03E0        ; DISALLOWED  # GREEK LETTER SAMPI
03E1        ; PVALID      # GREEK SMALL LETTER SAMPI
03E2        ; DISALLOWED  # COPTIC CAPITAL LETTER SHEI
03E3        ; PVALID      # COPTIC SMALL LETTER SHEI
03E4        ; DISALLOWED  # COPTIC CAPITAL LETTER FEI
03E5        ; PVALID      # COPTIC SMALL LETTER FEI
03E6        ; DISALLOWED  # COPTIC CAPITAL LETTER KHEI
03E7        ; PVALID      # COPTIC SMALL LETTER KHEI
03E8        ; DISALLOWED  # COPTIC CAPITAL LETTER HORI
03E9        ; PVALID      # COPTIC SMALL LETTER HORI
03EA        ; DISALLOWED  # COPTIC CAPITAL LETTER GANGIA
03EB        ; PVALID      # COPTIC SMALL LETTER GANGIA
03EC        ; DISALLOWED  # COPTIC CAPITAL LETTER SHIMA
03ED        ; PVALID      # COPTIC SMALL LETTER SHIMA
03EE        ; DISALLOWED  # COPTIC CAPITAL LETTER DEI
03EF        ; PVALID      # COPTIC SMALL LETTER DEI
03F0..03F2  ; DISALLOWED  # GREEK KAPPA SYMBOL..GREEK LUNATE SIGMA SYMBO
03F3        ; PVALID      # GREEK LETTER YOT
03F4..03F7  ; DISALLOWED  # GREEK CAPITAL THETA SYMBOL..GREEK CAPITAL LE
03F8        ; PVALID      # GREEK SMALL LETTER SHO
03F9..03FA  ; DISALLOWED  # GREEK CAPITAL LUNATE SIGMA SYMBOL..GREEK CAP
03FB..03FC  ; PVALID      # GREEK SMALL LETTER SAN..GREEK RHO WITH STROK
03FD..042F  ; DISALLOWED  # GREEK CAPITAL REVERSED LUNATE SIGMA SYMBOL..
0430..045F  ; PVALID      # CYRILLIC SMALL LETTER A..CYRILLIC SMALL LETT
0460        ; DISALLOWED  # CYRILLIC CAPITAL LETTER OMEGA
0461        ; PVALID      # CYRILLIC SMALL LETTER OMEGA
0462        ; DISALLOWED  # CYRILLIC CAPITAL LETTER YAT
0463        ; PVALID      # CYRILLIC SMALL LETTER YAT
0464        ; DISALLOWED  # CYRILLIC CAPITAL LETTER IOTIFIED E
0465        ; PVALID      # CYRILLIC SMALL LETTER IOTIFIED E
0466        ; DISALLOWED  # CYRILLIC CAPITAL LETTER LITTLE YUS
0467        ; PVALID      # CYRILLIC SMALL LETTER LITTLE YUS
0468        ; DISALLOWED  # CYRILLIC CAPITAL LETTER IOTIFIED LITTLE YUS
0469        ; PVALID      # CYRILLIC SMALL LETTER IOTIFIED LITTLE YUS
046A        ; DISALLOWED  # CYRILLIC CAPITAL LETTER BIG YUS
046B        ; PVALID      # CYRILLIC SMALL LETTER BIG YUS
046C        ; DISALLOWED  # CYRILLIC CAPITAL LETTER IOTIFIED BIG YUS
```

```
046D        ; PVALID      # CYRILLIC SMALL LETTER IOTIFIED BIG YUS
046E        ; DISALLOWED  # CYRILLIC CAPITAL LETTER KSI
046F        ; PVALID      # CYRILLIC SMALL LETTER KSI
0470        ; DISALLOWED  # CYRILLIC CAPITAL LETTER PSI
0471        ; PVALID      # CYRILLIC SMALL LETTER PSI
0472        ; DISALLOWED  # CYRILLIC CAPITAL LETTER FITA
0473        ; PVALID      # CYRILLIC SMALL LETTER FITA
0474        ; DISALLOWED  # CYRILLIC CAPITAL LETTER IZHITSA
0475        ; PVALID      # CYRILLIC SMALL LETTER IZHITSA
0476        ; DISALLOWED  # CYRILLIC CAPITAL LETTER IZHITSA WITH DOUBLE
0477        ; PVALID      # CYRILLIC SMALL LETTER IZHITSA WITH DOUBLE GR
0478        ; DISALLOWED  # CYRILLIC CAPITAL LETTER UK
0479        ; PVALID      # CYRILLIC SMALL LETTER UK
047A        ; DISALLOWED  # CYRILLIC CAPITAL LETTER ROUND OMEGA
047B        ; PVALID      # CYRILLIC SMALL LETTER ROUND OMEGA
047C        ; DISALLOWED  # CYRILLIC CAPITAL LETTER OMEGA WITH TITLO
047D        ; PVALID      # CYRILLIC SMALL LETTER OMEGA WITH TITLO
047E        ; DISALLOWED  # CYRILLIC CAPITAL LETTER OT
047F        ; PVALID      # CYRILLIC SMALL LETTER OT
0480        ; DISALLOWED  # CYRILLIC CAPITAL LETTER KOPPA
0481        ; PVALID      # CYRILLIC SMALL LETTER KOPPA
0482        ; DISALLOWED  # CYRILLIC THOUSANDS SIGN
0483..0487  ; PVALID      # COMBINING CYRILLIC TITLO..COMBINING CYRILLIC
0488..048A  ; DISALLOWED  # COMBINING CYRILLIC HUNDRED THOUSANDS SIGN..C
048B        ; PVALID      # CYRILLIC SMALL LETTER SHORT I WITH TAIL
048C        ; DISALLOWED  # CYRILLIC CAPITAL LETTER SEMISOFT SIGN
048D        ; PVALID      # CYRILLIC SMALL LETTER SEMISOFT SIGN
048E        ; DISALLOWED  # CYRILLIC CAPITAL LETTER ER WITH TICK
048F        ; PVALID      # CYRILLIC SMALL LETTER ER WITH TICK
0490        ; DISALLOWED  # CYRILLIC CAPITAL LETTER GHE WITH UPTURN
0491        ; PVALID      # CYRILLIC SMALL LETTER GHE WITH UPTURN
0492        ; DISALLOWED  # CYRILLIC CAPITAL LETTER GHE WITH STROKE
0493        ; PVALID      # CYRILLIC SMALL LETTER GHE WITH STROKE
0494        ; DISALLOWED  # CYRILLIC CAPITAL LETTER GHE WITH MIDDLE HOOK
0495        ; PVALID      # CYRILLIC SMALL LETTER GHE WITH MIDDLE HOOK
0496        ; DISALLOWED  # CYRILLIC CAPITAL LETTER ZHE WITH DESCENDER
0497        ; PVALID      # CYRILLIC SMALL LETTER ZHE WITH DESCENDER
0498        ; DISALLOWED  # CYRILLIC CAPITAL LETTER ZE WITH DESCENDER
0499        ; PVALID      # CYRILLIC SMALL LETTER ZE WITH DESCENDER
049A        ; DISALLOWED  # CYRILLIC CAPITAL LETTER KA WITH DESCENDER
049B        ; PVALID      # CYRILLIC SMALL LETTER KA WITH DESCENDER
049C        ; DISALLOWED  # CYRILLIC CAPITAL LETTER KA WITH VERTICAL STR
049D        ; PVALID      # CYRILLIC SMALL LETTER KA WITH VERTICAL STROK
049E        ; DISALLOWED  # CYRILLIC CAPITAL LETTER KA WITH STROKE
049F        ; PVALID      # CYRILLIC SMALL LETTER KA WITH STROKE
04A0        ; DISALLOWED  # CYRILLIC CAPITAL LETTER BASHKIR KA
04A1        ; PVALID      # CYRILLIC SMALL LETTER BASHKIR KA
04A2        ; DISALLOWED  # CYRILLIC CAPITAL LETTER EN WITH DESCENDER
```

```
04A3        ; PVALID      # CYRILLIC SMALL LETTER EN WITH DESCENDER
04A4        ; DISALLOWED  # CYRILLIC CAPITAL LIGATURE EN GHE
04A5        ; PVALID      # CYRILLIC SMALL LIGATURE EN GHE
04A6        ; DISALLOWED  # CYRILLIC CAPITAL LETTER PE WITH MIDDLE HOOK
04A7        ; PVALID      # CYRILLIC SMALL LETTER PE WITH MIDDLE HOOK
04A8        ; DISALLOWED  # CYRILLIC CAPITAL LETTER ABKHASIAN HA
04A9        ; PVALID      # CYRILLIC SMALL LETTER ABKHASIAN HA
04AA        ; DISALLOWED  # CYRILLIC CAPITAL LETTER ES WITH DESCENDER
04AB        ; PVALID      # CYRILLIC SMALL LETTER ES WITH DESCENDER
04AC        ; DISALLOWED  # CYRILLIC CAPITAL LETTER TE WITH DESCENDER
04AD        ; PVALID      # CYRILLIC SMALL LETTER TE WITH DESCENDER
04AE        ; DISALLOWED  # CYRILLIC CAPITAL LETTER STRAIGHT U
04AF        ; PVALID      # CYRILLIC SMALL LETTER STRAIGHT U
04B0        ; DISALLOWED  # CYRILLIC CAPITAL LETTER STRAIGHT U WITH STRO
04B1        ; PVALID      # CYRILLIC SMALL LETTER STRAIGHT U WITH STROKE
04B2        ; DISALLOWED  # CYRILLIC CAPITAL LETTER HA WITH DESCENDER
04B3        ; PVALID      # CYRILLIC SMALL LETTER HA WITH DESCENDER
04B4        ; DISALLOWED  # CYRILLIC CAPITAL LIGATURE TE TSE
04B5        ; PVALID      # CYRILLIC SMALL LIGATURE TE TSE
04B6        ; DISALLOWED  # CYRILLIC CAPITAL LETTER CHE WITH DESCENDER
04B7        ; PVALID      # CYRILLIC SMALL LETTER CHE WITH DESCENDER
04B8        ; DISALLOWED  # CYRILLIC CAPITAL LETTER CHE WITH VERTICAL ST
04B9        ; PVALID      # CYRILLIC SMALL LETTER CHE WITH VERTICAL STRO
04BA        ; DISALLOWED  # CYRILLIC CAPITAL LETTER SHHA
04BB        ; PVALID      # CYRILLIC SMALL LETTER SHHA
04BC        ; DISALLOWED  # CYRILLIC CAPITAL LETTER ABKHASIAN CHE
04BD        ; PVALID      # CYRILLIC SMALL LETTER ABKHASIAN CHE
04BE        ; DISALLOWED  # CYRILLIC CAPITAL LETTER ABKHASIAN CHE WITH D
04BF        ; PVALID      # CYRILLIC SMALL LETTER ABKHASIAN CHE WITH DES
04C0..04C1  ; DISALLOWED  # CYRILLIC LETTER PALOCHKA..CYRILLIC CAPITAL L
04C2        ; PVALID      # CYRILLIC SMALL LETTER ZHE WITH BREVE
04C3        ; DISALLOWED  # CYRILLIC CAPITAL LETTER KA WITH HOOK
04C4        ; PVALID      # CYRILLIC SMALL LETTER KA WITH HOOK
04C5        ; DISALLOWED  # CYRILLIC CAPITAL LETTER EL WITH TAIL
04C6        ; PVALID      # CYRILLIC SMALL LETTER EL WITH TAIL
04C7        ; DISALLOWED  # CYRILLIC CAPITAL LETTER EN WITH HOOK
04C8        ; PVALID      # CYRILLIC SMALL LETTER EN WITH HOOK
04C9        ; DISALLOWED  # CYRILLIC CAPITAL LETTER EN WITH TAIL
04CA        ; PVALID      # CYRILLIC SMALL LETTER EN WITH TAIL
04CB        ; DISALLOWED  # CYRILLIC CAPITAL LETTER KHAKASSIAN CHE
04CC        ; PVALID      # CYRILLIC SMALL LETTER KHAKASSIAN CHE
04CD        ; DISALLOWED  # CYRILLIC CAPITAL LETTER EM WITH TAIL
04CE..04CF  ; PVALID      # CYRILLIC SMALL LETTER EM WITH TAIL..CYRILLIC
04D0        ; DISALLOWED  # CYRILLIC CAPITAL LETTER A WITH BREVE
04D1        ; PVALID      # CYRILLIC SMALL LETTER A WITH BREVE
04D2        ; DISALLOWED  # CYRILLIC CAPITAL LETTER A WITH DIAERESIS
04D3        ; PVALID      # CYRILLIC SMALL LETTER A WITH DIAERESIS
04D4        ; DISALLOWED  # CYRILLIC CAPITAL LIGATURE A IE
```

```
04D5        ; PVALID      # CYRILLIC SMALL LIGATURE A IE
04D6        ; DISALLOWED  # CYRILLIC CAPITAL LETTER IE WITH BREVE
04D7        ; PVALID      # CYRILLIC SMALL LETTER IE WITH BREVE
04D8        ; DISALLOWED  # CYRILLIC CAPITAL LETTER SCHWA
04D9        ; PVALID      # CYRILLIC SMALL LETTER SCHWA
04DA        ; DISALLOWED  # CYRILLIC CAPITAL LETTER SCHWA WITH DIAERESIS
04DB        ; PVALID      # CYRILLIC SMALL LETTER SCHWA WITH DIAERESIS
04DC        ; DISALLOWED  # CYRILLIC CAPITAL LETTER ZHE WITH DIAERESIS
04DD        ; PVALID      # CYRILLIC SMALL LETTER ZHE WITH DIAERESIS
04DE        ; DISALLOWED  # CYRILLIC CAPITAL LETTER ZE WITH DIAERESIS
04DF        ; PVALID      # CYRILLIC SMALL LETTER ZE WITH DIAERESIS
04E0        ; DISALLOWED  # CYRILLIC CAPITAL LETTER ABKHASIAN DZE
04E1        ; PVALID      # CYRILLIC SMALL LETTER ABKHASIAN DZE
04E2        ; DISALLOWED  # CYRILLIC CAPITAL LETTER I WITH MACRON
04E3        ; PVALID      # CYRILLIC SMALL LETTER I WITH MACRON
04E4        ; DISALLOWED  # CYRILLIC CAPITAL LETTER I WITH DIAERESIS
04E5        ; PVALID      # CYRILLIC SMALL LETTER I WITH DIAERESIS
04E6        ; DISALLOWED  # CYRILLIC CAPITAL LETTER O WITH DIAERESIS
04E7        ; PVALID      # CYRILLIC SMALL LETTER O WITH DIAERESIS
04E8        ; DISALLOWED  # CYRILLIC CAPITAL LETTER BARRED O
04E9        ; PVALID      # CYRILLIC SMALL LETTER BARRED O
04EA        ; DISALLOWED  # CYRILLIC CAPITAL LETTER BARRED O WITH DIAERE
04EB        ; PVALID      # CYRILLIC SMALL LETTER BARRED O WITH DIAERESI
04EC        ; DISALLOWED  # CYRILLIC CAPITAL LETTER E WITH DIAERESIS
04ED        ; PVALID      # CYRILLIC SMALL LETTER E WITH DIAERESIS
04EE        ; DISALLOWED  # CYRILLIC CAPITAL LETTER U WITH MACRON
04EF        ; PVALID      # CYRILLIC SMALL LETTER U WITH MACRON
04F0        ; DISALLOWED  # CYRILLIC CAPITAL LETTER U WITH DIAERESIS
04F1        ; PVALID      # CYRILLIC SMALL LETTER U WITH DIAERESIS
04F2        ; DISALLOWED  # CYRILLIC CAPITAL LETTER U WITH DOUBLE ACUTE
04F3        ; PVALID      # CYRILLIC SMALL LETTER U WITH DOUBLE ACUTE
04F4        ; DISALLOWED  # CYRILLIC CAPITAL LETTER CHE WITH DIAERESIS
04F5        ; PVALID      # CYRILLIC SMALL LETTER CHE WITH DIAERESIS
04F6        ; DISALLOWED  # CYRILLIC CAPITAL LETTER GHE WITH DESCENDER
04F7        ; PVALID      # CYRILLIC SMALL LETTER GHE WITH DESCENDER
04F8        ; DISALLOWED  # CYRILLIC CAPITAL LETTER YERU WITH DIAERESIS
04F9        ; PVALID      # CYRILLIC SMALL LETTER YERU WITH DIAERESIS
04FA        ; DISALLOWED  # CYRILLIC CAPITAL LETTER GHE WITH STROKE AND
04FB        ; PVALID      # CYRILLIC SMALL LETTER GHE WITH STROKE AND HO
04FC        ; DISALLOWED  # CYRILLIC CAPITAL LETTER HA WITH HOOK
04FD        ; PVALID      # CYRILLIC SMALL LETTER HA WITH HOOK
04FE        ; DISALLOWED  # CYRILLIC CAPITAL LETTER HA WITH STROKE
04FF        ; PVALID      # CYRILLIC SMALL LETTER HA WITH STROKE
0500        ; DISALLOWED  # CYRILLIC CAPITAL LETTER KOMI DE
0501        ; PVALID      # CYRILLIC SMALL LETTER KOMI DE
0502        ; DISALLOWED  # CYRILLIC CAPITAL LETTER KOMI DJE
0503        ; PVALID      # CYRILLIC SMALL LETTER KOMI DJE
0504        ; DISALLOWED  # CYRILLIC CAPITAL LETTER KOMI ZJE
```

```
0505        ; PVALID      # CYRILLIC SMALL LETTER KOMI ZJE
0506        ; DISALLOWED  # CYRILLIC CAPITAL LETTER KOMI DZJE
0507        ; PVALID      # CYRILLIC SMALL LETTER KOMI DZJE
0508        ; DISALLOWED  # CYRILLIC CAPITAL LETTER KOMI LJE
0509        ; PVALID      # CYRILLIC SMALL LETTER KOMI LJE
050A        ; DISALLOWED  # CYRILLIC CAPITAL LETTER KOMI NJE
050B        ; PVALID      # CYRILLIC SMALL LETTER KOMI NJE
050C        ; DISALLOWED  # CYRILLIC CAPITAL LETTER KOMI SJE
050D        ; PVALID      # CYRILLIC SMALL LETTER KOMI SJE
050E        ; DISALLOWED  # CYRILLIC CAPITAL LETTER KOMI TJE
050F        ; PVALID      # CYRILLIC SMALL LETTER KOMI TJE
0510        ; DISALLOWED  # CYRILLIC CAPITAL LETTER REVERSED ZE
0511        ; PVALID      # CYRILLIC SMALL LETTER REVERSED ZE
0512        ; DISALLOWED  # CYRILLIC CAPITAL LETTER EL WITH HOOK
0513        ; PVALID      # CYRILLIC SMALL LETTER EL WITH HOOK
0514        ; DISALLOWED  # CYRILLIC CAPITAL LETTER LHA
0515        ; PVALID      # CYRILLIC SMALL LETTER LHA
0516        ; DISALLOWED  # CYRILLIC CAPITAL LETTER RHA
0517        ; PVALID      # CYRILLIC SMALL LETTER RHA
0518        ; DISALLOWED  # CYRILLIC CAPITAL LETTER YAE
0519        ; PVALID      # CYRILLIC SMALL LETTER YAE
051A        ; DISALLOWED  # CYRILLIC CAPITAL LETTER QA
051B        ; PVALID      # CYRILLIC SMALL LETTER QA
051C        ; DISALLOWED  # CYRILLIC CAPITAL LETTER WE
051D        ; PVALID      # CYRILLIC SMALL LETTER WE
051E        ; DISALLOWED  # CYRILLIC CAPITAL LETTER ALEUT KA
051F        ; PVALID      # CYRILLIC SMALL LETTER ALEUT KA
0520        ; DISALLOWED  # CYRILLIC CAPITAL LETTER EL WITH MIDDLE HOOK
0521        ; PVALID      # CYRILLIC SMALL LETTER EL WITH MIDDLE HOOK
0522        ; DISALLOWED  # CYRILLIC CAPITAL LETTER EN WITH MIDDLE HOOK
0523        ; PVALID      # CYRILLIC SMALL LETTER EN WITH MIDDLE HOOK
0524        ; DISALLOWED  # CYRILLIC CAPITAL LETTER PE WITH DESCENDER
0525        ; PVALID      # CYRILLIC SMALL LETTER PE WITH DESCENDER
0526        ; DISALLOWED  # CYRILLIC CAPITAL LETTER SHHA WITH DESCENDER
0527        ; PVALID      # CYRILLIC SMALL LETTER SHHA WITH DESCENDER
0528        ; DISALLOWED  # CYRILLIC CAPITAL LETTER EN WITH LEFT HOOK
0529        ; PVALID      # CYRILLIC SMALL LETTER EN WITH LEFT HOOK
052A        ; DISALLOWED  # CYRILLIC CAPITAL LETTER DZZHE
052B        ; PVALID      # CYRILLIC SMALL LETTER DZZHE
052C        ; DISALLOWED  # CYRILLIC CAPITAL LETTER DCHE
052D        ; PVALID      # CYRILLIC SMALL LETTER DCHE
052E        ; DISALLOWED  # CYRILLIC CAPITAL LETTER EL WITH DESCENDER
052F        ; PVALID      # CYRILLIC SMALL LETTER EL WITH DESCENDER
0530        ; UNASSIGNED  # <reserved>
0531..0556  ; DISALLOWED  # ARMENIAN CAPITAL LETTER AYB..ARMENIAN CAPITA
0557..0558  ; UNASSIGNED  # <reserved>..<reserved>
0559        ; PVALID      # ARMENIAN MODIFIER LETTER LEFT HALF RING
055A..055F  ; DISALLOWED  # ARMENIAN APOSTROPHE..ARMENIAN ABBREVIATION M
```

```
0560..0586  ; PVALID      # ARMENIAN SMALL LETTER TURNED AYB..ARMENIAN S
0587        ; DISALLOWED  # ARMENIAN SMALL LIGATURE ECH YIWN
0588        ; PVALID      # ARMENIAN SMALL LETTER YI WITH STROKE
0589..058A  ; DISALLOWED  # ARMENIAN FULL STOP..ARMENIAN HYPHEN
058B..058C  ; UNASSIGNED  # <reserved>..<reserved>
058D..058F  ; DISALLOWED  # RIGHT-FACING ARMENIAN ETERNITY SIGN..ARMENIA
0590        ; UNASSIGNED  # <reserved>
0591..05BD  ; PVALID      # HEBREW ACCENT ETNAHTA..HEBREW POINT METEG
05BE        ; DISALLOWED  # HEBREW PUNCTUATION MAQAF
05BF        ; PVALID      # HEBREW POINT RAFE
05C0        ; DISALLOWED  # HEBREW PUNCTUATION PASEQ
05C1..05C2  ; PVALID      # HEBREW POINT SHIN DOT..HEBREW POINT SIN DOT
05C3        ; DISALLOWED  # HEBREW PUNCTUATION SOF PASUQ
05C4..05C5  ; PVALID      # HEBREW MARK UPPER DOT..HEBREW MARK LOWER DOT
05C6        ; DISALLOWED  # HEBREW PUNCTUATION NUN HAFUKHA
05C7        ; PVALID      # HEBREW POINT QAMATS QATAN
05C8..05CF  ; UNASSIGNED  # <reserved>..<reserved>
05D0..05EA  ; PVALID      # HEBREW LETTER ALEF..HEBREW LETTER TAV
05EB..05EE  ; UNASSIGNED  # <reserved>..<reserved>
05EF..05F2  ; PVALID      # HEBREW YOD TRIANGLE..HEBREW LIGATURE YIDDISH
05F3..05F4  ; CONTEXTO    # HEBREW PUNCTUATION GERESH..HEBREW PUNCTUATIO
05F5..05FF  ; UNASSIGNED  # <reserved>..<reserved>
0600..060F  ; DISALLOWED  # ARABIC NUMBER SIGN..ARABIC SIGN MISRA
0610..061A  ; PVALID      # ARABIC SIGN SALLALLAHOU ALAYHE WASSALLAM..AR
061B..061C  ; DISALLOWED  # ARABIC SEMICOLON..ARABIC LETTER MARK
061D        ; UNASSIGNED  # <reserved>
061E..061F  ; DISALLOWED  # ARABIC TRIPLE DOT PUNCTUATION MARK..ARABIC Q
0620..063F  ; PVALID      # ARABIC LETTER KASHMIRI YEH..ARABIC LETTER FA
0640        ; DISALLOWED  # ARABIC TATWEEL
0641..065F  ; PVALID      # ARABIC LETTER FEH..ARABIC WAVY HAMZA BELOW
0660..0669  ; CONTEXTO    # ARABIC-INDIC DIGIT ZERO..ARABIC-INDIC DIGIT
066A..066D  ; DISALLOWED  # ARABIC PERCENT SIGN..ARABIC FIVE POINTED STA
066E..0674  ; PVALID      # ARABIC LETTER DOTLESS BEH..ARABIC LETTER HIG
0675..0678  ; DISALLOWED  # ARABIC LETTER HIGH HAMZA ALEF..ARABIC LETTER
0679..06D3  ; PVALID      # ARABIC LETTER TTEH..ARABIC LETTER YEH BARREE
06D4        ; DISALLOWED  # ARABIC FULL STOP
06D5..06DC  ; PVALID      # ARABIC LETTER AE..ARABIC SMALL HIGH SEEN
06DD..06DE  ; DISALLOWED  # ARABIC END OF AYAH..ARABIC START OF RUB EL H
06DF..06E8  ; PVALID      # ARABIC SMALL HIGH ROUNDED ZERO..ARABIC SMALL
06E9        ; DISALLOWED  # ARABIC PLACE OF SAJDAH
06EA..06EF  ; PVALID      # ARABIC EMPTY CENTRE LOW STOP..ARABIC LETTER
06F0..06F9  ; CONTEXTO    # EXTENDED ARABIC-INDIC DIGIT ZERO..EXTENDED A
06FA..06FF  ; PVALID      # ARABIC LETTER SHEEN WITH DOT BELOW..ARABIC L
0700..070D  ; DISALLOWED  # SYRIAC END OF PARAGRAPH..SYRIAC HARKLEAN AST
070E        ; UNASSIGNED  # <reserved>
070F        ; DISALLOWED  # SYRIAC ABBREVIATION MARK
0710..074A  ; PVALID      # SYRIAC LETTER ALAPH..SYRIAC BARREKH
074B..074C  ; UNASSIGNED  # <reserved>..<reserved>
```

```
074D..07B1  ; PVALID      # SYRIAC LETTER SOGDIAN ZHAIN..THAANA LETTER N
07B2..07BF  ; UNASSIGNED  # <reserved>..<reserved>
07C0..07F5  ; PVALID      # NKO DIGIT ZERO..NKO LOW TONE APOSTROPHE
07F6..07FA  ; DISALLOWED  # NKO SYMBOL OO DENNEN..NKO LAJANYALAN
07FB..07FC  ; UNASSIGNED  # <reserved>..<reserved>
07FD        ; PVALID      # NKO DANTAYALAN
07FE..07FF  ; DISALLOWED  # NKO DOROME SIGN..NKO TAMAN SIGN
0800..082D  ; PVALID      # SAMARITAN LETTER ALAF..SAMARITAN MARK NEQUDA
082E..082F  ; UNASSIGNED  # <reserved>..<reserved>
0830..083E  ; DISALLOWED  # SAMARITAN PUNCTUATION NEQUDAA..SAMARITAN PUN
083F        ; UNASSIGNED  # <reserved>
0840..085B  ; PVALID      # MANDAIC LETTER HALQA..MANDAIC GEMINATION MAR
085C..085D  ; UNASSIGNED  # <reserved>..<reserved>
085E        ; DISALLOWED  # MANDAIC PUNCTUATION
085F        ; UNASSIGNED  # <reserved>
0860..086A  ; PVALID      # SYRIAC LETTER MALAYALAM NGA..SYRIAC LETTER M
086B..089F  ; UNASSIGNED  # <reserved>..<reserved>
08A0..08B4  ; PVALID      # ARABIC LETTER BEH WITH SMALL V BELOW..ARABIC
08B5        ; UNASSIGNED  # <reserved>
08B6..08BD  ; PVALID      # ARABIC LETTER BEH WITH SMALL MEEM ABOVE..ARA
08BE..08D2  ; UNASSIGNED  # <reserved>..<reserved>
08D3..08E1  ; PVALID      # ARABIC SMALL LOW WAW..ARABIC SMALL HIGH SIGN
08E2        ; DISALLOWED  # ARABIC DISPUTED END OF AYAH
08E3..0957  ; PVALID      # ARABIC TURNED DAMMA BELOW..DEVANAGARI VOWEL
0958..095F  ; DISALLOWED  # DEVANAGARI LETTER QA..DEVANAGARI LETTER YYA
0960..0963  ; PVALID      # DEVANAGARI LETTER VOCALIC RR..DEVANAGARI VOW
0964..0965  ; DISALLOWED  # DEVANAGARI DANDA..DEVANAGARI DOUBLE DANDA
0966..096F  ; PVALID      # DEVANAGARI DIGIT ZERO..DEVANAGARI DIGIT NINE
0970        ; DISALLOWED  # DEVANAGARI ABBREVIATION SIGN
0971..0983  ; PVALID      # DEVANAGARI SIGN HIGH SPACING DOT..BENGALI SI
0984        ; UNASSIGNED  # <reserved>
0985..098C  ; PVALID      # BENGALI LETTER A..BENGALI LETTER VOCALIC L
098D..098E  ; UNASSIGNED  # <reserved>..<reserved>
098F..0990  ; PVALID      # BENGALI LETTER E..BENGALI LETTER AI
0991..0992  ; UNASSIGNED  # <reserved>..<reserved>
0993..09A8  ; PVALID      # BENGALI LETTER O..BENGALI LETTER NA
09A9        ; UNASSIGNED  # <reserved>
09AA..09B0  ; PVALID      # BENGALI LETTER PA..BENGALI LETTER RA
09B1        ; UNASSIGNED  # <reserved>
09B2        ; PVALID      # BENGALI LETTER LA
09B3..09B5  ; UNASSIGNED  # <reserved>..<reserved>
09B6..09B9  ; PVALID      # BENGALI LETTER SHA..BENGALI LETTER HA
09BA..09BB  ; UNASSIGNED  # <reserved>..<reserved>
09BC..09C4  ; PVALID      # BENGALI SIGN NUKTA..BENGALI VOWEL SIGN VOCAL
09C5..09C6  ; UNASSIGNED  # <reserved>..<reserved>
09C7..09C8  ; PVALID      # BENGALI VOWEL SIGN E..BENGALI VOWEL SIGN AI
09C9..09CA  ; UNASSIGNED  # <reserved>..<reserved>
09CB..09CE  ; PVALID      # BENGALI VOWEL SIGN O..BENGALI LETTER KHANDA
```

```
09CF..09D6  ; UNASSIGNED  # <reserved>..<reserved>
09D7        ; PVALID      # BENGALI AU LENGTH MARK
09D8..09DB  ; UNASSIGNED  # <reserved>..<reserved>
09DC..09DD  ; DISALLOWED  # BENGALI LETTER RRA..BENGALI LETTER RHA
09DE        ; UNASSIGNED  # <reserved>
09DF        ; DISALLOWED  # BENGALI LETTER YYA
09E0..09E3  ; PVALID      # BENGALI LETTER VOCALIC RR..BENGALI VOWEL SIG
09E4..09E5  ; UNASSIGNED  # <reserved>..<reserved>
09E6..09F1  ; PVALID      # BENGALI DIGIT ZERO..BENGALI LETTER RA WITH L
09F2..09FB  ; DISALLOWED  # BENGALI RUPEE MARK..BENGALI GANDA MARK
09FC        ; PVALID      # BENGALI LETTER VEDIC ANUSVARA
09FD        ; DISALLOWED  # BENGALI ABBREVIATION SIGN
09FE        ; PVALID      # BENGALI SANDHI MARK
09FF..0A00  ; UNASSIGNED  # <reserved>..<reserved>
0A01..0A03  ; PVALID      # GURMUKHI SIGN ADAK BINDI..GURMUKHI SIGN VISA
0A04        ; UNASSIGNED  # <reserved>
0A05..0A0A  ; PVALID      # GURMUKHI LETTER A..GURMUKHI LETTER UU
0A0B..0A0E  ; UNASSIGNED  # <reserved>..<reserved>
0A0F..0A10  ; PVALID      # GURMUKHI LETTER EE..GURMUKHI LETTER AI
0A11..0A12  ; UNASSIGNED  # <reserved>..<reserved>
0A13..0A28  ; PVALID      # GURMUKHI LETTER OO..GURMUKHI LETTER NA
0A29        ; UNASSIGNED  # <reserved>
0A2A..0A30  ; PVALID      # GURMUKHI LETTER PA..GURMUKHI LETTER RA
0A31        ; UNASSIGNED  # <reserved>
0A32        ; PVALID      # GURMUKHI LETTER LA
0A33        ; DISALLOWED  # GURMUKHI LETTER LLA
0A34        ; UNASSIGNED  # <reserved>
0A35        ; PVALID      # GURMUKHI LETTER VA
0A36        ; DISALLOWED  # GURMUKHI LETTER SHA
0A37        ; UNASSIGNED  # <reserved>
0A38..0A39  ; PVALID      # GURMUKHI LETTER SA..GURMUKHI LETTER HA
0A3A..0A3B  ; UNASSIGNED  # <reserved>..<reserved>
0A3C        ; PVALID      # GURMUKHI SIGN NUKTA
0A3D        ; UNASSIGNED  # <reserved>
0A3E..0A42  ; PVALID      # GURMUKHI VOWEL SIGN AA..GURMUKHI VOWEL SIGN
0A43..0A46  ; UNASSIGNED  # <reserved>..<reserved>
0A47..0A48  ; PVALID      # GURMUKHI VOWEL SIGN EE..GURMUKHI VOWEL SIGN
0A49..0A4A  ; UNASSIGNED  # <reserved>..<reserved>
0A4B..0A4D  ; PVALID      # GURMUKHI VOWEL SIGN OO..GURMUKHI SIGN VIRAMA
0A4E..0A50  ; UNASSIGNED  # <reserved>..<reserved>
0A51        ; PVALID      # GURMUKHI SIGN UDAAT
0A52..0A58  ; UNASSIGNED  # <reserved>..<reserved>
0A59..0A5B  ; DISALLOWED  # GURMUKHI LETTER KHHA..GURMUKHI LETTER ZA
0A5C        ; PVALID      # GURMUKHI LETTER RRA
0A5D        ; UNASSIGNED  # <reserved>
0A5E        ; DISALLOWED  # GURMUKHI LETTER FA
0A5F..0A65  ; UNASSIGNED  # <reserved>..<reserved>
0A66..0A75  ; PVALID      # GURMUKHI DIGIT ZERO..GURMUKHI SIGN YAKASH
```

```
0A76        ; DISALLOWED  # GURMUKHI ABBREVIATION SIGN
0A77..0A80  ; UNASSIGNED  # <reserved>..<reserved>
0A81..0A83  ; PVALID      # GUJARATI SIGN CANDRABINDU..GUJARATI SIGN VIS
0A84        ; UNASSIGNED  # <reserved>
0A85..0A8D  ; PVALID      # GUJARATI LETTER A..GUJARATI VOWEL CANDRA E
0A8E        ; UNASSIGNED  # <reserved>
0A8F..0A91  ; PVALID      # GUJARATI LETTER E..GUJARATI VOWEL CANDRA O
0A92        ; UNASSIGNED  # <reserved>
0A93..0AA8  ; PVALID      # GUJARATI LETTER O..GUJARATI LETTER NA
0AA9        ; UNASSIGNED  # <reserved>
0AAA..0AB0  ; PVALID      # GUJARATI LETTER PA..GUJARATI LETTER RA
0AB1        ; UNASSIGNED  # <reserved>
0AB2..0AB3  ; PVALID      # GUJARATI LETTER LA..GUJARATI LETTER LLA
0AB4        ; UNASSIGNED  # <reserved>
0AB5..0AB9  ; PVALID      # GUJARATI LETTER VA..GUJARATI LETTER HA
0ABA..0ABB  ; UNASSIGNED  # <reserved>..<reserved>
0ABC..0AC5  ; PVALID      # GUJARATI SIGN NUKTA..GUJARATI VOWEL SIGN CAN
0AC6        ; UNASSIGNED  # <reserved>
0AC7..0AC9  ; PVALID      # GUJARATI VOWEL SIGN E..GUJARATI VOWEL SIGN C
0ACA        ; UNASSIGNED  # <reserved>
0ACB..0ACD  ; PVALID      # GUJARATI VOWEL SIGN O..GUJARATI SIGN VIRAMA
0ACE..0ACF  ; UNASSIGNED  # <reserved>..<reserved>
0AD0        ; PVALID      # GUJARATI OM
0AD1..0ADF  ; UNASSIGNED  # <reserved>..<reserved>
0AE0..0AE3  ; PVALID      # GUJARATI LETTER VOCALIC RR..GUJARATI VOWEL S
0AE4..0AE5  ; UNASSIGNED  # <reserved>..<reserved>
0AE6..0AEF  ; PVALID      # GUJARATI DIGIT ZERO..GUJARATI DIGIT NINE
0AF0..0AF1  ; DISALLOWED  # GUJARATI ABBREVIATION SIGN..GUJARATI RUPEE S
0AF2..0AF8  ; UNASSIGNED  # <reserved>..<reserved>
0AF9..0AFF  ; PVALID      # GUJARATI LETTER ZHA..GUJARATI SIGN TWO-CIRCL
0B00        ; UNASSIGNED  # <reserved>
0B01..0B03  ; PVALID      # ORIYA SIGN CANDRABINDU..ORIYA SIGN VISARGA
0B04        ; UNASSIGNED  # <reserved>
0B05..0B0C  ; PVALID      # ORIYA LETTER A..ORIYA LETTER VOCALIC L
0B0D..0B0E  ; UNASSIGNED  # <reserved>..<reserved>
0B0F..0B10  ; PVALID      # ORIYA LETTER E..ORIYA LETTER AI
0B11..0B12  ; UNASSIGNED  # <reserved>..<reserved>
0B13..0B28  ; PVALID      # ORIYA LETTER O..ORIYA LETTER NA
0B29        ; UNASSIGNED  # <reserved>
0B2A..0B30  ; PVALID      # ORIYA LETTER PA..ORIYA LETTER RA
0B31        ; UNASSIGNED  # <reserved>
0B32..0B33  ; PVALID      # ORIYA LETTER LA..ORIYA LETTER LLA
0B34        ; UNASSIGNED  # <reserved>
0B35..0B39  ; PVALID      # ORIYA LETTER VA..ORIYA LETTER HA
0B3A..0B3B  ; UNASSIGNED  # <reserved>..<reserved>
0B3C..0B44  ; PVALID      # ORIYA SIGN NUKTA..ORIYA VOWEL SIGN VOCALIC R
0B45..0B46  ; UNASSIGNED  # <reserved>..<reserved>
0B47..0B48  ; PVALID      # ORIYA VOWEL SIGN E..ORIYA VOWEL SIGN AI
```

```
0B49..0B4A  ; UNASSIGNED  # <reserved>..<reserved>
0B4B..0B4D  ; PVALID      # ORIYA VOWEL SIGN O..ORIYA SIGN VIRAMA
0B4E..0B55  ; UNASSIGNED  # <reserved>..<reserved>
0B56..0B57  ; PVALID      # ORIYA AI LENGTH MARK..ORIYA AU LENGTH MARK
0B58..0B5B  ; UNASSIGNED  # <reserved>..<reserved>
0B5C..0B5D  ; DISALLOWED  # ORIYA LETTER RRA..ORIYA LETTER RHA
0B5E        ; UNASSIGNED  # <reserved>
0B5F..0B63  ; PVALID      # ORIYA LETTER YYA..ORIYA VOWEL SIGN VOCALIC L
0B64..0B65  ; UNASSIGNED  # <reserved>..<reserved>
0B66..0B6F  ; PVALID      # ORIYA DIGIT ZERO..ORIYA DIGIT NINE
0B70        ; DISALLOWED  # ORIYA ISSHAR
0B71        ; PVALID      # ORIYA LETTER WA
0B72..0B77  ; DISALLOWED  # ORIYA FRACTION ONE QUARTER..ORIYA FRACTION T
0B78..0B81  ; UNASSIGNED  # <reserved>..<reserved>
0B82..0B83  ; PVALID      # TAMIL SIGN ANUSVARA..TAMIL SIGN VISARGA
0B84        ; UNASSIGNED  # <reserved>
0B85..0B8A  ; PVALID      # TAMIL LETTER A..TAMIL LETTER UU
0B8B..0B8D  ; UNASSIGNED  # <reserved>..<reserved>
0B8E..0B90  ; PVALID      # TAMIL LETTER E..TAMIL LETTER AI
0B91        ; UNASSIGNED  # <reserved>
0B92..0B95  ; PVALID      # TAMIL LETTER O..TAMIL LETTER KA
0B96..0B98  ; UNASSIGNED  # <reserved>..<reserved>
0B99..0B9A  ; PVALID      # TAMIL LETTER NGA..TAMIL LETTER CA
0B9B        ; UNASSIGNED  # <reserved>
0B9C        ; PVALID      # TAMIL LETTER JA
0B9D        ; UNASSIGNED  # <reserved>
0B9E..0B9F  ; PVALID      # TAMIL LETTER NYA..TAMIL LETTER TTA
0BA0..0BA2  ; UNASSIGNED  # <reserved>..<reserved>
0BA3..0BA4  ; PVALID      # TAMIL LETTER NNA..TAMIL LETTER TA
0BA5..0BA7  ; UNASSIGNED  # <reserved>..<reserved>
0BA8..0BAA  ; PVALID      # TAMIL LETTER NA..TAMIL LETTER PA
0BAB..0BAD  ; UNASSIGNED  # <reserved>..<reserved>
0BAE..0BB9  ; PVALID      # TAMIL LETTER MA..TAMIL LETTER HA
0BBA..0BBD  ; UNASSIGNED  # <reserved>..<reserved>
0BBE..0BC2  ; PVALID      # TAMIL VOWEL SIGN AA..TAMIL VOWEL SIGN UU
0BC3..0BC5  ; UNASSIGNED  # <reserved>..<reserved>
0BC6..0BC8  ; PVALID      # TAMIL VOWEL SIGN E..TAMIL VOWEL SIGN AI
0BC9        ; UNASSIGNED  # <reserved>
0BCA..0BCD  ; PVALID      # TAMIL VOWEL SIGN O..TAMIL SIGN VIRAMA
0BCE..0BCF  ; UNASSIGNED  # <reserved>..<reserved>
0BD0        ; PVALID      # TAMIL OM
0BD1..0BD6  ; UNASSIGNED  # <reserved>..<reserved>
0BD7        ; PVALID      # TAMIL AU LENGTH MARK
0BD8..0BE5  ; UNASSIGNED  # <reserved>..<reserved>
0BE6..0BEF  ; PVALID      # TAMIL DIGIT ZERO..TAMIL DIGIT NINE
0BF0..0BFA  ; DISALLOWED  # TAMIL NUMBER TEN..TAMIL NUMBER SIGN
0BFB..0BFF  ; UNASSIGNED  # <reserved>..<reserved>
0C00..0C0C  ; PVALID      # TELUGU SIGN COMBINING CANDRABINDU ABOVE..TEL
```

```
0C0D        ; UNASSIGNED  # <reserved>
0C0E..0C10  ; PVALID      # TELUGU LETTER E..TELUGU LETTER AI
0C11        ; UNASSIGNED  # <reserved>
0C12..0C28  ; PVALID      # TELUGU LETTER O..TELUGU LETTER NA
0C29        ; UNASSIGNED  # <reserved>
0C2A..0C39  ; PVALID      # TELUGU LETTER PA..TELUGU LETTER HA
0C3A..0C3C  ; UNASSIGNED  # <reserved>..<reserved>
0C3D..0C44  ; PVALID      # TELUGU SIGN AVAGRAHA..TELUGU VOWEL SIGN VOCA
0C45        ; UNASSIGNED  # <reserved>
0C46..0C48  ; PVALID      # TELUGU VOWEL SIGN E..TELUGU VOWEL SIGN AI
0C49        ; UNASSIGNED  # <reserved>
0C4A..0C4D  ; PVALID      # TELUGU VOWEL SIGN O..TELUGU SIGN VIRAMA
0C4E..0C54  ; UNASSIGNED  # <reserved>..<reserved>
0C55..0C56  ; PVALID      # TELUGU LENGTH MARK..TELUGU AI LENGTH MARK
0C57        ; UNASSIGNED  # <reserved>
0C58..0C5A  ; PVALID      # TELUGU LETTER TSA..TELUGU LETTER RRRA
0C5B..0C5F  ; UNASSIGNED  # <reserved>..<reserved>
0C60..0C63  ; PVALID      # TELUGU LETTER VOCALIC RR..TELUGU VOWEL SIGN
0C64..0C65  ; UNASSIGNED  # <reserved>..<reserved>
0C66..0C6F  ; PVALID      # TELUGU DIGIT ZERO..TELUGU DIGIT NINE
0C70..0C77  ; UNASSIGNED  # <reserved>..<reserved>
0C78..0C7F  ; DISALLOWED  # TELUGU FRACTION DIGIT ZERO FOR ODD POWERS OF
0C80..0C83  ; PVALID      # KANNADA SIGN SPACING CANDRABINDU..KANNADA SI
0C84        ; DISALLOWED  # KANNADA SIGN SIDDHAM
0C85..0C8C  ; PVALID      # KANNADA LETTER A..KANNADA LETTER VOCALIC L
0C8D        ; UNASSIGNED  # <reserved>
0C8E..0C90  ; PVALID      # KANNADA LETTER E..KANNADA LETTER AI
0C91        ; UNASSIGNED  # <reserved>
0C92..0CA8  ; PVALID      # KANNADA LETTER O..KANNADA LETTER NA
0CA9        ; UNASSIGNED  # <reserved>
0CAA..0CB3  ; PVALID      # KANNADA LETTER PA..KANNADA LETTER LLA
0CB4        ; UNASSIGNED  # <reserved>
0CB5..0CB9  ; PVALID      # KANNADA LETTER VA..KANNADA LETTER HA
0CBA..0CBB  ; UNASSIGNED  # <reserved>..<reserved>
0CBC..0CC4  ; PVALID      # KANNADA SIGN NUKTA..KANNADA VOWEL SIGN VOCAL
0CC5        ; UNASSIGNED  # <reserved>
0CC6..0CC8  ; PVALID      # KANNADA VOWEL SIGN E..KANNADA VOWEL SIGN AI
0CC9        ; UNASSIGNED  # <reserved>
0CCA..0CCD  ; PVALID      # KANNADA VOWEL SIGN O..KANNADA SIGN VIRAMA
0CCE..0CD4  ; UNASSIGNED  # <reserved>..<reserved>
0CD5..0CD6  ; PVALID      # KANNADA LENGTH MARK..KANNADA AI LENGTH MARK
0CD7..0CDD  ; UNASSIGNED  # <reserved>..<reserved>
0CDE        ; PVALID      # KANNADA LETTER FA
0CDF        ; UNASSIGNED  # <reserved>
0CE0..0CE3  ; PVALID      # KANNADA LETTER VOCALIC RR..KANNADA VOWEL SIG
0CE4..0CE5  ; UNASSIGNED  # <reserved>..<reserved>
0CE6..0CEF  ; PVALID      # KANNADA DIGIT ZERO..KANNADA DIGIT NINE
0CF0        ; UNASSIGNED  # <reserved>
```

```
0CF1..0CF2  ; PVALID      # KANNADA SIGN JIHVAMULIYA..KANNADA SIGN UPADH
0CF3..0CFF  ; UNASSIGNED  # <reserved>..<reserved>
0D00..0D03  ; PVALID      # MALAYALAM SIGN COMBINING ANUSVARA ABOVE..MAL
0D04        ; UNASSIGNED  # <reserved>
0D05..0D0C  ; PVALID      # MALAYALAM LETTER A..MALAYALAM LETTER VOCALIC
0D0D        ; UNASSIGNED  # <reserved>
0D0E..0D10  ; PVALID      # MALAYALAM LETTER E..MALAYALAM LETTER AI
0D11        ; UNASSIGNED  # <reserved>
0D12..0D44  ; PVALID      # MALAYALAM LETTER O..MALAYALAM VOWEL SIGN VOC
0D45        ; UNASSIGNED  # <reserved>
0D46..0D48  ; PVALID      # MALAYALAM VOWEL SIGN E..MALAYALAM VOWEL SIGN
0D49        ; UNASSIGNED  # <reserved>
0D4A..0D4E  ; PVALID      # MALAYALAM VOWEL SIGN O..MALAYALAM LETTER DOT
0D4F        ; DISALLOWED  # MALAYALAM SIGN PARA
0D50..0D53  ; UNASSIGNED  # <reserved>..<reserved>
0D54..0D57  ; PVALID      # MALAYALAM LETTER CHILLU M..MALAYALAM AU LENG
0D58..0D5E  ; DISALLOWED  # MALAYALAM FRACTION ONE ONE-HUNDRED-AND-SIXTI
0D5F..0D63  ; PVALID      # MALAYALAM LETTER ARCHAIC II..MALAYALAM VOWEL
0D64..0D65  ; UNASSIGNED  # <reserved>..<reserved>
0D66..0D6F  ; PVALID      # MALAYALAM DIGIT ZERO..MALAYALAM DIGIT NINE
0D70..0D79  ; DISALLOWED  # MALAYALAM NUMBER TEN..MALAYALAM DATE MARK
0D7A..0D7F  ; PVALID      # MALAYALAM LETTER CHILLU NN..MALAYALAM LETTER
0D80..0D81  ; UNASSIGNED  # <reserved>..<reserved>
0D82..0D83  ; PVALID      # SINHALA SIGN ANUSVARAYA..SINHALA SIGN VISARG
0D84        ; UNASSIGNED  # <reserved>
0D85..0D96  ; PVALID      # SINHALA LETTER AYANNA..SINHALA LETTER AUYANN
0D97..0D99  ; UNASSIGNED  # <reserved>..<reserved>
0D9A..0DB1  ; PVALID      # SINHALA LETTER ALPAPRAANA KAYANNA..SINHALA L
0DB2        ; UNASSIGNED  # <reserved>
0DB3..0DBB  ; PVALID      # SINHALA LETTER SANYAKA DAYANNA..SINHALA LETT
0DBC        ; UNASSIGNED  # <reserved>
0DBD        ; PVALID      # SINHALA LETTER DANTAJA LAYANNA
0DBE..0DBF  ; UNASSIGNED  # <reserved>..<reserved>
0DC0..0DC6  ; PVALID      # SINHALA LETTER VAYANNA..SINHALA LETTER FAYAN
0DC7..0DC9  ; UNASSIGNED  # <reserved>..<reserved>
0DCA        ; PVALID      # SINHALA SIGN AL-LAKUNA
0DCB..0DCE  ; UNASSIGNED  # <reserved>..<reserved>
0DCF..0DD4  ; PVALID      # SINHALA VOWEL SIGN AELA-PILLA..SINHALA VOWEL
0DD5        ; UNASSIGNED  # <reserved>
0DD6        ; PVALID      # SINHALA VOWEL SIGN DIGA PAA-PILLA
0DD7        ; UNASSIGNED  # <reserved>
0DD8..0DDF  ; PVALID      # SINHALA VOWEL SIGN GAETTA-PILLA..SINHALA VOW
0DE0..0DE5  ; UNASSIGNED  # <reserved>..<reserved>
0DE6..0DEF  ; PVALID      # SINHALA LITH DIGIT ZERO..SINHALA LITH DIGIT
0DF0..0DF1  ; UNASSIGNED  # <reserved>..<reserved>
0DF2..0DF3  ; PVALID      # SINHALA VOWEL SIGN DIGA GAETTA-PILLA..SINHAL
0DF4        ; DISALLOWED  # SINHALA PUNCTUATION KUNDDALIYA
0DF5..0E00  ; UNASSIGNED  # <reserved>..<reserved>
```

```
0E01..0E32  ; PVALID      # THAI CHARACTER KO KAI..THAI CHARACTER SARA A
0E33        ; DISALLOWED  # THAI CHARACTER SARA AM
0E34..0E3A  ; PVALID      # THAI CHARACTER SARA I..THAI CHARACTER PHINTH
0E3B..0E3E  ; UNASSIGNED  # <reserved>..<reserved>
0E3F        ; DISALLOWED  # THAI CURRENCY SYMBOL BAHT
0E40..0E4E  ; PVALID      # THAI CHARACTER SARA E..THAI CHARACTER YAMAKK
0E4F        ; DISALLOWED  # THAI CHARACTER FONGMAN
0E50..0E59  ; PVALID      # THAI DIGIT ZERO..THAI DIGIT NINE
0E5A..0E5B  ; DISALLOWED  # THAI CHARACTER ANGKHANKHU..THAI CHARACTER KH
0E5C..0E80  ; UNASSIGNED  # <reserved>..<reserved>
0E81..0E82  ; PVALID      # LAO LETTER KO..LAO LETTER KHO SUNG
0E83        ; UNASSIGNED  # <reserved>
0E84        ; PVALID      # LAO LETTER KHO TAM
0E85..0E86  ; UNASSIGNED  # <reserved>..<reserved>
0E87..0E88  ; PVALID      # LAO LETTER NGO..LAO LETTER CO
0E89        ; UNASSIGNED  # <reserved>
0E8A        ; PVALID      # LAO LETTER SO TAM
0E8B..0E8C  ; UNASSIGNED  # <reserved>..<reserved>
0E8D        ; PVALID      # LAO LETTER NYO
0E8E..0E93  ; UNASSIGNED  # <reserved>..<reserved>
0E94..0E97  ; PVALID      # LAO LETTER DO..LAO LETTER THO TAM
0E98        ; UNASSIGNED  # <reserved>
0E99..0E9F  ; PVALID      # LAO LETTER NO..LAO LETTER FO SUNG
0EA0        ; UNASSIGNED  # <reserved>
0EA1..0EA3  ; PVALID      # LAO LETTER MO..LAO LETTER LO LING
0EA4        ; UNASSIGNED  # <reserved>
0EA5        ; PVALID      # LAO LETTER LO LOOT
0EA6        ; UNASSIGNED  # <reserved>
0EA7        ; PVALID      # LAO LETTER WO
0EA8..0EA9  ; UNASSIGNED  # <reserved>..<reserved>
0EAA..0EAB  ; PVALID      # LAO LETTER SO SUNG..LAO LETTER HO SUNG
0EAC        ; UNASSIGNED  # <reserved>
0EAD..0EB2  ; PVALID      # LAO LETTER O..LAO VOWEL SIGN AA
0EB3        ; DISALLOWED  # LAO VOWEL SIGN AM
0EB4..0EB9  ; PVALID      # LAO VOWEL SIGN I..LAO VOWEL SIGN UU
0EBA        ; UNASSIGNED  # <reserved>
0EBB..0EBD  ; PVALID      # LAO VOWEL SIGN MAI KON..LAO SEMIVOWEL SIGN N
0EBE..0EBF  ; UNASSIGNED  # <reserved>..<reserved>
0EC0..0EC4  ; PVALID      # LAO VOWEL SIGN E..LAO VOWEL SIGN AI
0EC5        ; UNASSIGNED  # <reserved>
0EC6        ; PVALID      # LAO KO LA
0EC7        ; UNASSIGNED  # <reserved>
0EC8..0ECD  ; PVALID      # LAO TONE MAI EK..LAO NIGGAHITA
0ECE..0ECF  ; UNASSIGNED  # <reserved>..<reserved>
0ED0..0ED9  ; PVALID      # LAO DIGIT ZERO..LAO DIGIT NINE
0EDA..0EDB  ; UNASSIGNED  # <reserved>..<reserved>
0EDC..0EDD  ; DISALLOWED  # LAO HO NO..LAO HO MO
0EDE..0EDF  ; PVALID      # LAO LETTER KHMU GO..LAO LETTER KHMU NYO
```

```
0EE0..0EFF  ; UNASSIGNED  # <reserved>..<reserved>
0F00        ; PVALID      # TIBETAN SYLLABLE OM
0F01..0F0A  ; DISALLOWED  # TIBETAN MARK GTER YIG MGO TRUNCATED A..TIBET
0F0B        ; PVALID      # TIBETAN MARK INTERSYLLABIC TSHEG
0F0C..0F17  ; DISALLOWED  # TIBETAN MARK DELIMITER TSHEG BSTAR..TIBETAN
0F18..0F19  ; PVALID      # TIBETAN ASTROLOGICAL SIGN -KHYUD PA..TIBETAN
0F1A..0F1F  ; DISALLOWED  # TIBETAN SIGN RDEL DKAR GCIG..TIBETAN SIGN RD
0F20..0F29  ; PVALID      # TIBETAN DIGIT ZERO..TIBETAN DIGIT NINE
0F2A..0F34  ; DISALLOWED  # TIBETAN DIGIT HALF ONE..TIBETAN MARK BSDUS R
0F35        ; PVALID      # TIBETAN MARK NGAS BZUNG NYI ZLA
0F36        ; DISALLOWED  # TIBETAN MARK CARET -DZUD RTAGS BZHI MIG CAN
0F37        ; PVALID      # TIBETAN MARK NGAS BZUNG SGOR RTAGS
0F38        ; DISALLOWED  # TIBETAN MARK CHE MGO
0F39        ; PVALID      # TIBETAN MARK TSA -PHRU
0F3A..0F3D  ; DISALLOWED  # TIBETAN MARK GUG RTAGS GYON..TIBETAN MARK AN
0F3E..0F42  ; PVALID      # TIBETAN SIGN YAR TSHES..TIBETAN LETTER GA
0F43        ; DISALLOWED  # TIBETAN LETTER GHA
0F44..0F47  ; PVALID      # TIBETAN LETTER NGA..TIBETAN LETTER JA
0F48        ; UNASSIGNED  # <reserved>
0F49..0F4C  ; PVALID      # TIBETAN LETTER NYA..TIBETAN LETTER DDA
0F4D        ; DISALLOWED  # TIBETAN LETTER DDHA
0F4E..0F51  ; PVALID      # TIBETAN LETTER NNA..TIBETAN LETTER DA
0F52        ; DISALLOWED  # TIBETAN LETTER DHA
0F53..0F56  ; PVALID      # TIBETAN LETTER NA..TIBETAN LETTER BA
0F57        ; DISALLOWED  # TIBETAN LETTER BHA
0F58..0F5B  ; PVALID      # TIBETAN LETTER MA..TIBETAN LETTER DZA
0F5C        ; DISALLOWED  # TIBETAN LETTER DZHA
0F5D..0F68  ; PVALID      # TIBETAN LETTER WA..TIBETAN LETTER A
0F69        ; DISALLOWED  # TIBETAN LETTER KSSA
0F6A..0F6C  ; PVALID      # TIBETAN LETTER FIXED-FORM RA..TIBETAN LETTER
0F6D..0F70  ; UNASSIGNED  # <reserved>..<reserved>
0F71..0F72  ; PVALID      # TIBETAN VOWEL SIGN AA..TIBETAN VOWEL SIGN I
0F73        ; DISALLOWED  # TIBETAN VOWEL SIGN II
0F74        ; PVALID      # TIBETAN VOWEL SIGN U
0F75..0F79  ; DISALLOWED  # TIBETAN VOWEL SIGN UU..TIBETAN VOWEL SIGN VO
0F7A..0F80  ; PVALID      # TIBETAN VOWEL SIGN E..TIBETAN VOWEL SIGN REV
0F81        ; DISALLOWED  # TIBETAN VOWEL SIGN REVERSED II
0F82..0F84  ; PVALID      # TIBETAN SIGN NYI ZLA NAA DA..TIBETAN MARK HA
0F85        ; DISALLOWED  # TIBETAN MARK PALUTA
0F86..0F92  ; PVALID      # TIBETAN SIGN LCI RTAGS..TIBETAN SUBJOINED LE
0F93        ; DISALLOWED  # TIBETAN SUBJOINED LETTER GHA
0F94..0F97  ; PVALID      # TIBETAN SUBJOINED LETTER NGA..TIBETAN SUBJOI
0F98        ; UNASSIGNED  # <reserved>
0F99..0F9C  ; PVALID      # TIBETAN SUBJOINED LETTER NYA..TIBETAN SUBJOI
0F9D        ; DISALLOWED  # TIBETAN SUBJOINED LETTER DDHA
0F9E..0FA1  ; PVALID      # TIBETAN SUBJOINED LETTER NNA..TIBETAN SUBJOI
0FA2        ; DISALLOWED  # TIBETAN SUBJOINED LETTER DHA
0FA3..0FA6  ; PVALID      # TIBETAN SUBJOINED LETTER NA..TIBETAN SUBJOIN
```

```
0FA7        ; DISALLOWED  # TIBETAN SUBJOINED LETTER BHA
0FA8..0FAB  ; PVALID      # TIBETAN SUBJOINED LETTER MA..TIBETAN SUBJOIN
0FAC        ; DISALLOWED  # TIBETAN SUBJOINED LETTER DZHA
0FAD..0FB8  ; PVALID      # TIBETAN SUBJOINED LETTER WA..TIBETAN SUBJOIN
0FB9        ; DISALLOWED  # TIBETAN SUBJOINED LETTER KSSA
0FBA..0FBC  ; PVALID      # TIBETAN SUBJOINED LETTER FIXED-FORM WA..TIBE
0FBD        ; UNASSIGNED  # <reserved>
0FBE..0FC5  ; DISALLOWED  # TIBETAN KU RU KHA..TIBETAN SYMBOL RDO RJE
0FC6        ; PVALID      # TIBETAN SYMBOL PADMA GDAN
0FC7..0FCC  ; DISALLOWED  # TIBETAN SYMBOL RDO RJE RGYA GRAM..TIBETAN SY
0FCD        ; UNASSIGNED  # <reserved>
0FCE..0FDA  ; DISALLOWED  # TIBETAN SIGN RDEL NAG RDEL DKAR..TIBETAN MAR
0FDB..0FFF  ; UNASSIGNED  # <reserved>..<reserved>
1000..1049  ; PVALID      # MYANMAR LETTER KA..MYANMAR DIGIT NINE
104A..104F  ; DISALLOWED  # MYANMAR SIGN LITTLE SECTION..MYANMAR SYMBOL
1050..109D  ; PVALID      # MYANMAR LETTER SHA..MYANMAR VOWEL SIGN AITON
109E..10C5  ; DISALLOWED  # MYANMAR SYMBOL SHAN ONE..GEORGIAN CAPITAL LE
10C6        ; UNASSIGNED  # <reserved>
10C7        ; DISALLOWED  # GEORGIAN CAPITAL LETTER YN
10C8..10CC  ; UNASSIGNED  # <reserved>..<reserved>
10CD        ; DISALLOWED  # GEORGIAN CAPITAL LETTER AEN
10CE..10CF  ; UNASSIGNED  # <reserved>..<reserved>
10D0..10FA  ; PVALID      # GEORGIAN LETTER AN..GEORGIAN LETTER AIN
10FB..10FC  ; DISALLOWED  # GEORGIAN PARAGRAPH SEPARATOR..MODIFIER LETTE
10FD..10FF  ; PVALID      # GEORGIAN LETTER AEN..GEORGIAN LETTER LABIAL
1100..11FF  ; DISALLOWED  # HANGUL CHOSEONG KIYEOK..HANGUL JONGSEONG SSA
1200..1248  ; PVALID      # ETHIOPIC SYLLABLE HA..ETHIOPIC SYLLABLE QWA
1249        ; UNASSIGNED  # <reserved>
124A..124D  ; PVALID      # ETHIOPIC SYLLABLE QWI..ETHIOPIC SYLLABLE QWE
124E..124F  ; UNASSIGNED  # <reserved>..<reserved>
1250..1256  ; PVALID      # ETHIOPIC SYLLABLE QHA..ETHIOPIC SYLLABLE QHO
1257        ; UNASSIGNED  # <reserved>
1258        ; PVALID      # ETHIOPIC SYLLABLE QHWA
1259        ; UNASSIGNED  # <reserved>
125A..125D  ; PVALID      # ETHIOPIC SYLLABLE QHWI..ETHIOPIC SYLLABLE QH
125E..125F  ; UNASSIGNED  # <reserved>..<reserved>
1260..1288  ; PVALID      # ETHIOPIC SYLLABLE BA..ETHIOPIC SYLLABLE XWA
1289        ; UNASSIGNED  # <reserved>
128A..128D  ; PVALID      # ETHIOPIC SYLLABLE XWI..ETHIOPIC SYLLABLE XWE
128E..128F  ; UNASSIGNED  # <reserved>..<reserved>
1290..12B0  ; PVALID      # ETHIOPIC SYLLABLE NA..ETHIOPIC SYLLABLE KWA
12B1        ; UNASSIGNED  # <reserved>
12B2..12B5  ; PVALID      # ETHIOPIC SYLLABLE KWI..ETHIOPIC SYLLABLE KWE
12B6..12B7  ; UNASSIGNED  # <reserved>..<reserved>
12B8..12BE  ; PVALID      # ETHIOPIC SYLLABLE KXA..ETHIOPIC SYLLABLE KXO
12BF        ; UNASSIGNED  # <reserved>
12C0        ; PVALID      # ETHIOPIC SYLLABLE KXWA
12C1        ; UNASSIGNED  # <reserved>
```

```
12C2..12C5  ; PVALID      # ETHIOPIC SYLLABLE KXWI..ETHIOPIC SYLLABLE KX
12C6..12C7  ; UNASSIGNED  # <reserved>..<reserved>
12C8..12D6  ; PVALID      # ETHIOPIC SYLLABLE WA..ETHIOPIC SYLLABLE PHAR
12D7        ; UNASSIGNED  # <reserved>
12D8..1310  ; PVALID      # ETHIOPIC SYLLABLE ZA..ETHIOPIC SYLLABLE GWA
1311        ; UNASSIGNED  # <reserved>
1312..1315  ; PVALID      # ETHIOPIC SYLLABLE GWI..ETHIOPIC SYLLABLE GWE
1316..1317  ; UNASSIGNED  # <reserved>..<reserved>
1318..135A  ; PVALID      # ETHIOPIC SYLLABLE GGA..ETHIOPIC SYLLABLE FYA
135B..135C  ; UNASSIGNED  # <reserved>..<reserved>
135D..135F  ; PVALID      # ETHIOPIC COMBINING GEMINATION AND VOWEL LENG
1360..137C  ; DISALLOWED  # ETHIOPIC SECTION MARK..ETHIOPIC NUMBER TEN T
137D..137F  ; UNASSIGNED  # <reserved>..<reserved>
1380..138F  ; PVALID      # ETHIOPIC SYLLABLE SEBATBEIT MWA..ETHIOPIC SY
1390..1399  ; DISALLOWED  # ETHIOPIC TONAL MARK YIZET..ETHIOPIC TONAL MA
139A..139F  ; UNASSIGNED  # <reserved>..<reserved>
13A0..13F5  ; PVALID      # CHEROKEE LETTER A..CHEROKEE LETTER MV
13F6..13F7  ; UNASSIGNED  # <reserved>..<reserved>
13F8..13FD  ; DISALLOWED  # CHEROKEE SMALL LETTER YE..CHEROKEE SMALL LET
13FE..13FF  ; UNASSIGNED  # <reserved>..<reserved>
1400        ; DISALLOWED  # CANADIAN SYLLABICS HYPHEN
1401..166C  ; PVALID      # CANADIAN SYLLABICS E..CANADIAN SYLLABICS CAR
166D..166E  ; DISALLOWED  # CANADIAN SYLLABICS CHI SIGN..CANADIAN SYLLAB
166F..167F  ; PVALID      # CANADIAN SYLLABICS QAI..CANADIAN SYLLABICS B
1680        ; DISALLOWED  # OGHAM SPACE MARK
1681..169A  ; PVALID      # OGHAM LETTER BEITH..OGHAM LETTER PEITH
169B..169C  ; DISALLOWED  # OGHAM FEATHER MARK..OGHAM REVERSED FEATHER M
169D..169F  ; UNASSIGNED  # <reserved>..<reserved>
16A0..16EA  ; PVALID      # RUNIC LETTER FEHU FEOH FE F..RUNIC LETTER X
16EB..16F0  ; DISALLOWED  # RUNIC SINGLE PUNCTUATION..RUNIC BELGTHOR SYM
16F1..16F8  ; PVALID      # RUNIC LETTER K..RUNIC LETTER FRANKS CASKET A
16F9..16FF  ; UNASSIGNED  # <reserved>..<reserved>
1700..170C  ; PVALID      # TAGALOG LETTER A..TAGALOG LETTER YA
170D        ; UNASSIGNED  # <reserved>
170E..1714  ; PVALID      # TAGALOG LETTER LA..TAGALOG SIGN VIRAMA
1715..171F  ; UNASSIGNED  # <reserved>..<reserved>
1720..1734  ; PVALID      # HANUNOO LETTER A..HANUNOO SIGN PAMUDPOD
1735..1736  ; DISALLOWED  # PHILIPPINE SINGLE PUNCTUATION..PHILIPPINE DO
1737..173F  ; UNASSIGNED  # <reserved>..<reserved>
1740..1753  ; PVALID      # BUHID LETTER A..BUHID VOWEL SIGN U
1754..175F  ; UNASSIGNED  # <reserved>..<reserved>
1760..176C  ; PVALID      # TAGBANWA LETTER A..TAGBANWA LETTER YA
176D        ; UNASSIGNED  # <reserved>
176E..1770  ; PVALID      # TAGBANWA LETTER LA..TAGBANWA LETTER SA
1771        ; UNASSIGNED  # <reserved>
1772..1773  ; PVALID      # TAGBANWA VOWEL SIGN I..TAGBANWA VOWEL SIGN U
1774..177F  ; UNASSIGNED  # <reserved>..<reserved>
1780..17B3  ; PVALID      # KHMER LETTER KA..KHMER INDEPENDENT VOWEL QAU
```

```
17B4..17B5  ; DISALLOWED  # KHMER VOWEL INHERENT AQ..KHMER VOWEL INHEREN
17B6..17D3  ; PVALID      # KHMER VOWEL SIGN AA..KHMER SIGN BATHAMASAT
17D4..17D6  ; DISALLOWED  # KHMER SIGN KHAN..KHMER SIGN CAMNUC PII KUUH
17D7        ; PVALID      # KHMER SIGN LEK TOO
17D8..17DB  ; DISALLOWED  # KHMER SIGN BEYYAL..KHMER CURRENCY SYMBOL RIE
17DC..17DD  ; PVALID      # KHMER SIGN AVAKRAHASANYA..KHMER SIGN ATTHACA
17DE..17DF  ; UNASSIGNED  # <reserved>..<reserved>
17E0..17E9  ; PVALID      # KHMER DIGIT ZERO..KHMER DIGIT NINE
17EA..17EF  ; UNASSIGNED  # <reserved>..<reserved>
17F0..17F9  ; DISALLOWED  # KHMER SYMBOL LEK ATTAK SON..KHMER SYMBOL LEK
17FA..17FF  ; UNASSIGNED  # <reserved>..<reserved>
1800..180E  ; DISALLOWED  # MONGOLIAN BIRGA..MONGOLIAN VOWEL SEPARATOR
180F        ; UNASSIGNED  # <reserved>
1810..1819  ; PVALID      # MONGOLIAN DIGIT ZERO..MONGOLIAN DIGIT NINE
181A..181F  ; UNASSIGNED  # <reserved>..<reserved>
1820..1878  ; PVALID      # MONGOLIAN LETTER A..MONGOLIAN LETTER CHA WIT
1879..187F  ; UNASSIGNED  # <reserved>..<reserved>
1880..18AA  ; PVALID      # MONGOLIAN LETTER ALI GALI ANUSVARA ONE..MONG
18AB..18AF  ; UNASSIGNED  # <reserved>..<reserved>
18B0..18F5  ; PVALID      # CANADIAN SYLLABICS OY..CANADIAN SYLLABICS CA
18F6..18FF  ; UNASSIGNED  # <reserved>..<reserved>
1900..191E  ; PVALID      # LIMBU VOWEL-CARRIER LETTER..LIMBU LETTER TRA
191F        ; UNASSIGNED  # <reserved>
1920..192B  ; PVALID      # LIMBU VOWEL SIGN A..LIMBU SUBJOINED LETTER W
192C..192F  ; UNASSIGNED  # <reserved>..<reserved>
1930..193B  ; PVALID      # LIMBU SMALL LETTER KA..LIMBU SIGN SA-I
193C..193F  ; UNASSIGNED  # <reserved>..<reserved>
1940        ; DISALLOWED  # LIMBU SIGN LOO
1941..1943  ; UNASSIGNED  # <reserved>..<reserved>
1944..1945  ; DISALLOWED  # LIMBU EXCLAMATION MARK..LIMBU QUESTION MARK
1946..196D  ; PVALID      # LIMBU DIGIT ZERO..TAI LE LETTER AI
196E..196F  ; UNASSIGNED  # <reserved>..<reserved>
1970..1974  ; PVALID      # TAI LE LETTER TONE-2..TAI LE LETTER TONE-6
1975..197F  ; UNASSIGNED  # <reserved>..<reserved>
1980..19AB  ; PVALID      # NEW TAI LUE LETTER HIGH QA..NEW TAI LUE LETT
19AC..19AF  ; UNASSIGNED  # <reserved>..<reserved>
19B0..19C9  ; PVALID      # NEW TAI LUE VOWEL SIGN VOWEL SHORTENER..NEW
19CA..19CF  ; UNASSIGNED  # <reserved>..<reserved>
19D0..19D9  ; PVALID      # NEW TAI LUE DIGIT ZERO..NEW TAI LUE DIGIT NI
19DA        ; DISALLOWED  # NEW TAI LUE THAM DIGIT ONE
19DB..19DD  ; UNASSIGNED  # <reserved>..<reserved>
19DE..19FF  ; DISALLOWED  # NEW TAI LUE SIGN LAE..KHMER SYMBOL DAP-PRAM
1A00..1A1B  ; PVALID      # BUGINESE LETTER KA..BUGINESE VOWEL SIGN AE
1A1C..1A1D  ; UNASSIGNED  # <reserved>..<reserved>
1A1E..1A1F  ; DISALLOWED  # BUGINESE PALLAWA..BUGINESE END OF SECTION
1A20..1A5E  ; PVALID      # TAI THAM LETTER HIGH KA..TAI THAM CONSONANT
1A5F        ; UNASSIGNED  # <reserved>
1A60..1A7C  ; PVALID      # TAI THAM SIGN SAKOT..TAI THAM SIGN KHUEN-LUE
```

```
1A7D..1A7E  ; UNASSIGNED  # <reserved>..<reserved>
1A7F..1A89  ; PVALID      # TAI THAM COMBINING CRYPTOGRAMMIC DOT..TAI TH
1A8A..1A8F  ; UNASSIGNED  # <reserved>..<reserved>
1A90..1A99  ; PVALID      # TAI THAM THAM DIGIT ZERO..TAI THAM THAM DIGI
1A9A..1A9F  ; UNASSIGNED  # <reserved>..<reserved>
1AA0..1AA6  ; DISALLOWED  # TAI THAM SIGN WIANG..TAI THAM SIGN REVERSED
1AA7        ; PVALID      # TAI THAM SIGN MAI YAMOK
1AA8..1AAD  ; DISALLOWED  # TAI THAM SIGN KAAN..TAI THAM SIGN CAANG
1AAE..1AAF  ; UNASSIGNED  # <reserved>..<reserved>
1AB0..1ABD  ; PVALID      # COMBINING DOUBLED CIRCUMFLEX ACCENT..COMBINI
1ABE        ; DISALLOWED  # COMBINING PARENTHESES OVERLAY
1ABF..1AFF  ; UNASSIGNED  # <reserved>..<reserved>
1B00..1B4B  ; PVALID      # BALINESE SIGN ULU RICEM..BALINESE LETTER ASY
1B4C..1B4F  ; UNASSIGNED  # <reserved>..<reserved>
1B50..1B59  ; PVALID      # BALINESE DIGIT ZERO..BALINESE DIGIT NINE
1B5A..1B6A  ; DISALLOWED  # BALINESE PANTI..BALINESE MUSICAL SYMBOL DANG
1B6B..1B73  ; PVALID      # BALINESE MUSICAL SYMBOL COMBINING TEGEH..BAL
1B74..1B7C  ; DISALLOWED  # BALINESE MUSICAL SYMBOL RIGHT-HAND OPEN DUG.
1B7D..1B7F  ; UNASSIGNED  # <reserved>..<reserved>
1B80..1BF3  ; PVALID      # SUNDANESE SIGN PANYECEK..BATAK PANONGONAN
1BF4..1BFB  ; UNASSIGNED  # <reserved>..<reserved>
1BFC..1BFF  ; DISALLOWED  # BATAK SYMBOL BINDU NA METEK..BATAK SYMBOL BI
1C00..1C37  ; PVALID      # LEPCHA LETTER KA..LEPCHA SIGN NUKTA
1C38..1C3A  ; UNASSIGNED  # <reserved>..<reserved>
1C3B..1C3F  ; DISALLOWED  # LEPCHA PUNCTUATION TA-ROL..LEPCHA PUNCTUATIO
1C40..1C49  ; PVALID      # LEPCHA DIGIT ZERO..LEPCHA DIGIT NINE
1C4A..1C4C  ; UNASSIGNED  # <reserved>..<reserved>
1C4D..1C7D  ; PVALID      # LEPCHA LETTER TTA..OL CHIKI AHAD
1C7E..1C88  ; DISALLOWED  # OL CHIKI PUNCTUATION MUCAAD..CYRILLIC SMALL
1C89..1C8F  ; UNASSIGNED  # <reserved>..<reserved>
1C90..1CBA  ; DISALLOWED  # GEORGIAN MTAVRULI CAPITAL LETTER AN..GEORGIA
1CBB..1CBC  ; UNASSIGNED  # <reserved>..<reserved>
1CBD..1CC7  ; DISALLOWED  # GEORGIAN MTAVRULI CAPITAL LETTER AEN..SUNDAN
1CC8..1CCF  ; UNASSIGNED  # <reserved>..<reserved>
1CD0..1CD2  ; PVALID      # VEDIC TONE KARSHANA..VEDIC TONE PRENKHA
1CD3        ; DISALLOWED  # VEDIC SIGN NIHSHVASA
1CD4..1CF9  ; PVALID      # VEDIC SIGN YAJURVEDIC MIDLINE SVARITA..VEDIC
1CFA..1CFF  ; UNASSIGNED  # <reserved>..<reserved>
1D00..1D2B  ; PVALID      # LATIN LETTER SMALL CAPITAL A..CYRILLIC LETTE
1D2C..1D2E  ; DISALLOWED  # MODIFIER LETTER CAPITAL A..MODIFIER LETTER C
1D2F        ; PVALID      # MODIFIER LETTER CAPITAL BARRED B
1D30..1D3A  ; DISALLOWED  # MODIFIER LETTER CAPITAL D..MODIFIER LETTER C
1D3B        ; PVALID      # MODIFIER LETTER CAPITAL REVERSED N
1D3C..1D4D  ; DISALLOWED  # MODIFIER LETTER CAPITAL O..MODIFIER LETTER S
1D4E        ; PVALID      # MODIFIER LETTER SMALL TURNED I
1D4F..1D6A  ; DISALLOWED  # MODIFIER LETTER SMALL K..GREEK SUBSCRIPT SMA
1D6B..1D77  ; PVALID      # LATIN SMALL LETTER UE..LATIN SMALL LETTER TU
1D78        ; DISALLOWED  # MODIFIER LETTER CYRILLIC EN
```

```
1D79..1D9A  ; PVALID      # LATIN SMALL LETTER INSULAR G..LATIN SMALL LE
1D9B..1DBF  ; DISALLOWED  # MODIFIER LETTER SMALL TURNED ALPHA..MODIFIER
1DC0..1DF9  ; PVALID      # COMBINING DOTTED GRAVE ACCENT..COMBINING WID
1DFA        ; UNASSIGNED  # <reserved>
1DFB..1DFF  ; PVALID      # COMBINING DELETION MARK..COMBINING RIGHT ARR
1E00        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH RING BELOW
1E01        ; PVALID      # LATIN SMALL LETTER A WITH RING BELOW
1E02        ; DISALLOWED  # LATIN CAPITAL LETTER B WITH DOT ABOVE
1E03        ; PVALID      # LATIN SMALL LETTER B WITH DOT ABOVE
1E04        ; DISALLOWED  # LATIN CAPITAL LETTER B WITH DOT BELOW
1E05        ; PVALID      # LATIN SMALL LETTER B WITH DOT BELOW
1E06        ; DISALLOWED  # LATIN CAPITAL LETTER B WITH LINE BELOW
1E07        ; PVALID      # LATIN SMALL LETTER B WITH LINE BELOW
1E08        ; DISALLOWED  # LATIN CAPITAL LETTER C WITH CEDILLA AND ACUT
1E09        ; PVALID      # LATIN SMALL LETTER C WITH CEDILLA AND ACUTE
1E0A        ; DISALLOWED  # LATIN CAPITAL LETTER D WITH DOT ABOVE
1E0B        ; PVALID      # LATIN SMALL LETTER D WITH DOT ABOVE
1E0C        ; DISALLOWED  # LATIN CAPITAL LETTER D WITH DOT BELOW
1E0D        ; PVALID      # LATIN SMALL LETTER D WITH DOT BELOW
1E0E        ; DISALLOWED  # LATIN CAPITAL LETTER D WITH LINE BELOW
1E0F        ; PVALID      # LATIN SMALL LETTER D WITH LINE BELOW
1E10        ; DISALLOWED  # LATIN CAPITAL LETTER D WITH CEDILLA
1E11        ; PVALID      # LATIN SMALL LETTER D WITH CEDILLA
1E12        ; DISALLOWED  # LATIN CAPITAL LETTER D WITH CIRCUMFLEX BELOW
1E13        ; PVALID      # LATIN SMALL LETTER D WITH CIRCUMFLEX BELOW
1E14        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH MACRON AND GRAVE
1E15        ; PVALID      # LATIN SMALL LETTER E WITH MACRON AND GRAVE
1E16        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH MACRON AND ACUTE
1E17        ; PVALID      # LATIN SMALL LETTER E WITH MACRON AND ACUTE
1E18        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH CIRCUMFLEX BELOW
1E19        ; PVALID      # LATIN SMALL LETTER E WITH CIRCUMFLEX BELOW
1E1A        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH TILDE BELOW
1E1B        ; PVALID      # LATIN SMALL LETTER E WITH TILDE BELOW
1E1C        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH CEDILLA AND BREV
1E1D        ; PVALID      # LATIN SMALL LETTER E WITH CEDILLA AND BREVE
1E1E        ; DISALLOWED  # LATIN CAPITAL LETTER F WITH DOT ABOVE
1E1F        ; PVALID      # LATIN SMALL LETTER F WITH DOT ABOVE
1E20        ; DISALLOWED  # LATIN CAPITAL LETTER G WITH MACRON
1E21        ; PVALID      # LATIN SMALL LETTER G WITH MACRON
1E22        ; DISALLOWED  # LATIN CAPITAL LETTER H WITH DOT ABOVE
1E23        ; PVALID      # LATIN SMALL LETTER H WITH DOT ABOVE
1E24        ; DISALLOWED  # LATIN CAPITAL LETTER H WITH DOT BELOW
1E25        ; PVALID      # LATIN SMALL LETTER H WITH DOT BELOW
1E26        ; DISALLOWED  # LATIN CAPITAL LETTER H WITH DIAERESIS
1E27        ; PVALID      # LATIN SMALL LETTER H WITH DIAERESIS
1E28        ; DISALLOWED  # LATIN CAPITAL LETTER H WITH CEDILLA
1E29        ; PVALID      # LATIN SMALL LETTER H WITH CEDILLA
1E2A        ; DISALLOWED  # LATIN CAPITAL LETTER H WITH BREVE BELOW
```

```
1E2B        ; PVALID      # LATIN SMALL LETTER H WITH BREVE BELOW
1E2C        ; DISALLOWED  # LATIN CAPITAL LETTER I WITH TILDE BELOW
1E2D        ; PVALID      # LATIN SMALL LETTER I WITH TILDE BELOW
1E2E        ; DISALLOWED  # LATIN CAPITAL LETTER I WITH DIAERESIS AND AC
1E2F        ; PVALID      # LATIN SMALL LETTER I WITH DIAERESIS AND ACUT
1E30        ; DISALLOWED  # LATIN CAPITAL LETTER K WITH ACUTE
1E31        ; PVALID      # LATIN SMALL LETTER K WITH ACUTE
1E32        ; DISALLOWED  # LATIN CAPITAL LETTER K WITH DOT BELOW
1E33        ; PVALID      # LATIN SMALL LETTER K WITH DOT BELOW
1E34        ; DISALLOWED  # LATIN CAPITAL LETTER K WITH LINE BELOW
1E35        ; PVALID      # LATIN SMALL LETTER K WITH LINE BELOW
1E36        ; DISALLOWED  # LATIN CAPITAL LETTER L WITH DOT BELOW
1E37        ; PVALID      # LATIN SMALL LETTER L WITH DOT BELOW
1E38        ; DISALLOWED  # LATIN CAPITAL LETTER L WITH DOT BELOW AND MA
1E39        ; PVALID      # LATIN SMALL LETTER L WITH DOT BELOW AND MACR
1E3A        ; DISALLOWED  # LATIN CAPITAL LETTER L WITH LINE BELOW
1E3B        ; PVALID      # LATIN SMALL LETTER L WITH LINE BELOW
1E3C        ; DISALLOWED  # LATIN CAPITAL LETTER L WITH CIRCUMFLEX BELOW
1E3D        ; PVALID      # LATIN SMALL LETTER L WITH CIRCUMFLEX BELOW
1E3E        ; DISALLOWED  # LATIN CAPITAL LETTER M WITH ACUTE
1E3F        ; PVALID      # LATIN SMALL LETTER M WITH ACUTE
1E40        ; DISALLOWED  # LATIN CAPITAL LETTER M WITH DOT ABOVE
1E41        ; PVALID      # LATIN SMALL LETTER M WITH DOT ABOVE
1E42        ; DISALLOWED  # LATIN CAPITAL LETTER M WITH DOT BELOW
1E43        ; PVALID      # LATIN SMALL LETTER M WITH DOT BELOW
1E44        ; DISALLOWED  # LATIN CAPITAL LETTER N WITH DOT ABOVE
1E45        ; PVALID      # LATIN SMALL LETTER N WITH DOT ABOVE
1E46        ; DISALLOWED  # LATIN CAPITAL LETTER N WITH DOT BELOW
1E47        ; PVALID      # LATIN SMALL LETTER N WITH DOT BELOW
1E48        ; DISALLOWED  # LATIN CAPITAL LETTER N WITH LINE BELOW
1E49        ; PVALID      # LATIN SMALL LETTER N WITH LINE BELOW
1E4A        ; DISALLOWED  # LATIN CAPITAL LETTER N WITH CIRCUMFLEX BELOW
1E4B        ; PVALID      # LATIN SMALL LETTER N WITH CIRCUMFLEX BELOW
1E4C        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH TILDE AND ACUTE
1E4D        ; PVALID      # LATIN SMALL LETTER O WITH TILDE AND ACUTE
1E4E        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH TILDE AND DIAERE
1E4F        ; PVALID      # LATIN SMALL LETTER O WITH TILDE AND DIAERESI
1E50        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH MACRON AND GRAVE
1E51        ; PVALID      # LATIN SMALL LETTER O WITH MACRON AND GRAVE
1E52        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH MACRON AND ACUTE
1E53        ; PVALID      # LATIN SMALL LETTER O WITH MACRON AND ACUTE
1E54        ; DISALLOWED  # LATIN CAPITAL LETTER P WITH ACUTE
1E55        ; PVALID      # LATIN SMALL LETTER P WITH ACUTE
1E56        ; DISALLOWED  # LATIN CAPITAL LETTER P WITH DOT ABOVE
1E57        ; PVALID      # LATIN SMALL LETTER P WITH DOT ABOVE
1E58        ; DISALLOWED  # LATIN CAPITAL LETTER R WITH DOT ABOVE
1E59        ; PVALID      # LATIN SMALL LETTER R WITH DOT ABOVE
1E5A        ; DISALLOWED  # LATIN CAPITAL LETTER R WITH DOT BELOW
```

```
1E5B        ; PVALID      # LATIN SMALL LETTER R WITH DOT BELOW
1E5C        ; DISALLOWED  # LATIN CAPITAL LETTER R WITH DOT BELOW AND MA
1E5D        ; PVALID      # LATIN SMALL LETTER R WITH DOT BELOW AND MACR
1E5E        ; DISALLOWED  # LATIN CAPITAL LETTER R WITH LINE BELOW
1E5F        ; PVALID      # LATIN SMALL LETTER R WITH LINE BELOW
1E60        ; DISALLOWED  # LATIN CAPITAL LETTER S WITH DOT ABOVE
1E61        ; PVALID      # LATIN SMALL LETTER S WITH DOT ABOVE
1E62        ; DISALLOWED  # LATIN CAPITAL LETTER S WITH DOT BELOW
1E63        ; PVALID      # LATIN SMALL LETTER S WITH DOT BELOW
1E64        ; DISALLOWED  # LATIN CAPITAL LETTER S WITH ACUTE AND DOT AB
1E65        ; PVALID      # LATIN SMALL LETTER S WITH ACUTE AND DOT ABOV
1E66        ; DISALLOWED  # LATIN CAPITAL LETTER S WITH CARON AND DOT AB
1E67        ; PVALID      # LATIN SMALL LETTER S WITH CARON AND DOT ABOV
1E68        ; DISALLOWED  # LATIN CAPITAL LETTER S WITH DOT BELOW AND DO
1E69        ; PVALID      # LATIN SMALL LETTER S WITH DOT BELOW AND DOT
1E6A        ; DISALLOWED  # LATIN CAPITAL LETTER T WITH DOT ABOVE
1E6B        ; PVALID      # LATIN SMALL LETTER T WITH DOT ABOVE
1E6C        ; DISALLOWED  # LATIN CAPITAL LETTER T WITH DOT BELOW
1E6D        ; PVALID      # LATIN SMALL LETTER T WITH DOT BELOW
1E6E        ; DISALLOWED  # LATIN CAPITAL LETTER T WITH LINE BELOW
1E6F        ; PVALID      # LATIN SMALL LETTER T WITH LINE BELOW
1E70        ; DISALLOWED  # LATIN CAPITAL LETTER T WITH CIRCUMFLEX BELOW
1E71        ; PVALID      # LATIN SMALL LETTER T WITH CIRCUMFLEX BELOW
1E72        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH DIAERESIS BELOW
1E73        ; PVALID      # LATIN SMALL LETTER U WITH DIAERESIS BELOW
1E74        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH TILDE BELOW
1E75        ; PVALID      # LATIN SMALL LETTER U WITH TILDE BELOW
1E76        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH CIRCUMFLEX BELOW
1E77        ; PVALID      # LATIN SMALL LETTER U WITH CIRCUMFLEX BELOW
1E78        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH TILDE AND ACUTE
1E79        ; PVALID      # LATIN SMALL LETTER U WITH TILDE AND ACUTE
1E7A        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH MACRON AND DIAER
1E7B        ; PVALID      # LATIN SMALL LETTER U WITH MACRON AND DIAERES
1E7C        ; DISALLOWED  # LATIN CAPITAL LETTER V WITH TILDE
1E7D        ; PVALID      # LATIN SMALL LETTER V WITH TILDE
1E7E        ; DISALLOWED  # LATIN CAPITAL LETTER V WITH DOT BELOW
1E7F        ; PVALID      # LATIN SMALL LETTER V WITH DOT BELOW
1E80        ; DISALLOWED  # LATIN CAPITAL LETTER W WITH GRAVE
1E81        ; PVALID      # LATIN SMALL LETTER W WITH GRAVE
1E82        ; DISALLOWED  # LATIN CAPITAL LETTER W WITH ACUTE
1E83        ; PVALID      # LATIN SMALL LETTER W WITH ACUTE
1E84        ; DISALLOWED  # LATIN CAPITAL LETTER W WITH DIAERESIS
1E85        ; PVALID      # LATIN SMALL LETTER W WITH DIAERESIS
1E86        ; DISALLOWED  # LATIN CAPITAL LETTER W WITH DOT ABOVE
1E87        ; PVALID      # LATIN SMALL LETTER W WITH DOT ABOVE
1E88        ; DISALLOWED  # LATIN CAPITAL LETTER W WITH DOT BELOW
1E89        ; PVALID      # LATIN SMALL LETTER W WITH DOT BELOW
1E8A        ; DISALLOWED  # LATIN CAPITAL LETTER X WITH DOT ABOVE
```

```
1E8B        ; PVALID      # LATIN SMALL LETTER X WITH DOT ABOVE
1E8C        ; DISALLOWED  # LATIN CAPITAL LETTER X WITH DIAERESIS
1E8D        ; PVALID      # LATIN SMALL LETTER X WITH DIAERESIS
1E8E        ; DISALLOWED  # LATIN CAPITAL LETTER Y WITH DOT ABOVE
1E8F        ; PVALID      # LATIN SMALL LETTER Y WITH DOT ABOVE
1E90        ; DISALLOWED  # LATIN CAPITAL LETTER Z WITH CIRCUMFLEX
1E91        ; PVALID      # LATIN SMALL LETTER Z WITH CIRCUMFLEX
1E92        ; DISALLOWED  # LATIN CAPITAL LETTER Z WITH DOT BELOW
1E93        ; PVALID      # LATIN SMALL LETTER Z WITH DOT BELOW
1E94        ; DISALLOWED  # LATIN CAPITAL LETTER Z WITH LINE BELOW
1E95..1E99  ; PVALID      # LATIN SMALL LETTER Z WITH LINE BELOW..LATIN
1E9A..1E9B  ; DISALLOWED  # LATIN SMALL LETTER A WITH RIGHT HALF RING..L
1E9C..1E9D  ; PVALID      # LATIN SMALL LETTER LONG S WITH DIAGONAL STRO
1E9E        ; DISALLOWED  # LATIN CAPITAL LETTER SHARP S
1E9F        ; PVALID      # LATIN SMALL LETTER DELTA
1EA0        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH DOT BELOW
1EA1        ; PVALID      # LATIN SMALL LETTER A WITH DOT BELOW
1EA2        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH HOOK ABOVE
1EA3        ; PVALID      # LATIN SMALL LETTER A WITH HOOK ABOVE
1EA4        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH CIRCUMFLEX AND A
1EA5        ; PVALID      # LATIN SMALL LETTER A WITH CIRCUMFLEX AND ACU
1EA6        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH CIRCUMFLEX AND G
1EA7        ; PVALID      # LATIN SMALL LETTER A WITH CIRCUMFLEX AND GRA
1EA8        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH CIRCUMFLEX AND H
1EA9        ; PVALID      # LATIN SMALL LETTER A WITH CIRCUMFLEX AND HOO
1EAA        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH CIRCUMFLEX AND T
1EAB        ; PVALID      # LATIN SMALL LETTER A WITH CIRCUMFLEX AND TIL
1EAC        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH CIRCUMFLEX AND D
1EAD        ; PVALID      # LATIN SMALL LETTER A WITH CIRCUMFLEX AND DOT
1EAE        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH BREVE AND ACUTE
1EAF        ; PVALID      # LATIN SMALL LETTER A WITH BREVE AND ACUTE
1EB0        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH BREVE AND GRAVE
1EB1        ; PVALID      # LATIN SMALL LETTER A WITH BREVE AND GRAVE
1EB2        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH BREVE AND HOOK A
1EB3        ; PVALID      # LATIN SMALL LETTER A WITH BREVE AND HOOK ABO
1EB4        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH BREVE AND TILDE
1EB5        ; PVALID      # LATIN SMALL LETTER A WITH BREVE AND TILDE
1EB6        ; DISALLOWED  # LATIN CAPITAL LETTER A WITH BREVE AND DOT BE
1EB7        ; PVALID      # LATIN SMALL LETTER A WITH BREVE AND DOT BELO
1EB8        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH DOT BELOW
1EB9        ; PVALID      # LATIN SMALL LETTER E WITH DOT BELOW
1EBA        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH HOOK ABOVE
1EBB        ; PVALID      # LATIN SMALL LETTER E WITH HOOK ABOVE
1EBC        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH TILDE
1EBD        ; PVALID      # LATIN SMALL LETTER E WITH TILDE
1EBE        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH CIRCUMFLEX AND A
1EBF        ; PVALID      # LATIN SMALL LETTER E WITH CIRCUMFLEX AND ACU
1EC0        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH CIRCUMFLEX AND G
```

```
1EC1        ; PVALID      # LATIN SMALL LETTER E WITH CIRCUMFLEX AND GRA
1EC2        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH CIRCUMFLEX AND H
1EC3        ; PVALID      # LATIN SMALL LETTER E WITH CIRCUMFLEX AND HOO
1EC4        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH CIRCUMFLEX AND T
1EC5        ; PVALID      # LATIN SMALL LETTER E WITH CIRCUMFLEX AND TIL
1EC6        ; DISALLOWED  # LATIN CAPITAL LETTER E WITH CIRCUMFLEX AND D
1EC7        ; PVALID      # LATIN SMALL LETTER E WITH CIRCUMFLEX AND DOT
1EC8        ; DISALLOWED  # LATIN CAPITAL LETTER I WITH HOOK ABOVE
1EC9        ; PVALID      # LATIN SMALL LETTER I WITH HOOK ABOVE
1ECA        ; DISALLOWED  # LATIN CAPITAL LETTER I WITH DOT BELOW
1ECB        ; PVALID      # LATIN SMALL LETTER I WITH DOT BELOW
1ECC        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH DOT BELOW
1ECD        ; PVALID      # LATIN SMALL LETTER O WITH DOT BELOW
1ECE        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH HOOK ABOVE
1ECF        ; PVALID      # LATIN SMALL LETTER O WITH HOOK ABOVE
1ED0        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH CIRCUMFLEX AND A
1ED1        ; PVALID      # LATIN SMALL LETTER O WITH CIRCUMFLEX AND ACU
1ED2        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH CIRCUMFLEX AND G
1ED3        ; PVALID      # LATIN SMALL LETTER O WITH CIRCUMFLEX AND GRA
1ED4        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH CIRCUMFLEX AND H
1ED5        ; PVALID      # LATIN SMALL LETTER O WITH CIRCUMFLEX AND HOO
1ED6        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH CIRCUMFLEX AND T
1ED7        ; PVALID      # LATIN SMALL LETTER O WITH CIRCUMFLEX AND TIL
1ED8        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH CIRCUMFLEX AND D
1ED9        ; PVALID      # LATIN SMALL LETTER O WITH CIRCUMFLEX AND DOT
1EDA        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH HORN AND ACUTE
1EDB        ; PVALID      # LATIN SMALL LETTER O WITH HORN AND ACUTE
1EDC        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH HORN AND GRAVE
1EDD        ; PVALID      # LATIN SMALL LETTER O WITH HORN AND GRAVE
1EDE        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH HORN AND HOOK AB
1EDF        ; PVALID      # LATIN SMALL LETTER O WITH HORN AND HOOK ABOV
1EE0        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH HORN AND TILDE
1EE1        ; PVALID      # LATIN SMALL LETTER O WITH HORN AND TILDE
1EE2        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH HORN AND DOT BEL
1EE3        ; PVALID      # LATIN SMALL LETTER O WITH HORN AND DOT BELOW
1EE4        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH DOT BELOW
1EE5        ; PVALID      # LATIN SMALL LETTER U WITH DOT BELOW
1EE6        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH HOOK ABOVE
1EE7        ; PVALID      # LATIN SMALL LETTER U WITH HOOK ABOVE
1EE8        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH HORN AND ACUTE
1EE9        ; PVALID      # LATIN SMALL LETTER U WITH HORN AND ACUTE
1EEA        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH HORN AND GRAVE
1EEB        ; PVALID      # LATIN SMALL LETTER U WITH HORN AND GRAVE
1EEC        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH HORN AND HOOK AB
1EED        ; PVALID      # LATIN SMALL LETTER U WITH HORN AND HOOK ABOV
1EEE        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH HORN AND TILDE
1EEF        ; PVALID      # LATIN SMALL LETTER U WITH HORN AND TILDE
1EF0        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH HORN AND DOT BEL
```

```
1EF1        ; PVALID      # LATIN SMALL LETTER U WITH HORN AND DOT BELOW
1EF2        ; DISALLOWED  # LATIN CAPITAL LETTER Y WITH GRAVE
1EF3        ; PVALID      # LATIN SMALL LETTER Y WITH GRAVE
1EF4        ; DISALLOWED  # LATIN CAPITAL LETTER Y WITH DOT BELOW
1EF5        ; PVALID      # LATIN SMALL LETTER Y WITH DOT BELOW
1EF6        ; DISALLOWED  # LATIN CAPITAL LETTER Y WITH HOOK ABOVE
1EF7        ; PVALID      # LATIN SMALL LETTER Y WITH HOOK ABOVE
1EF8        ; DISALLOWED  # LATIN CAPITAL LETTER Y WITH TILDE
1EF9        ; PVALID      # LATIN SMALL LETTER Y WITH TILDE
1EFA        ; DISALLOWED  # LATIN CAPITAL LETTER MIDDLE-WELSH LL
1EFB        ; PVALID      # LATIN SMALL LETTER MIDDLE-WELSH LL
1EFC        ; DISALLOWED  # LATIN CAPITAL LETTER MIDDLE-WELSH V
1EFD        ; PVALID      # LATIN SMALL LETTER MIDDLE-WELSH V
1EFE        ; DISALLOWED  # LATIN CAPITAL LETTER Y WITH LOOP
1EFF..1F07  ; PVALID      # LATIN SMALL LETTER Y WITH LOOP..GREEK SMALL
1F08..1F0F  ; DISALLOWED  # GREEK CAPITAL LETTER ALPHA WITH PSILI..GREEK
1F10..1F15  ; PVALID      # GREEK SMALL LETTER EPSILON WITH PSILI..GREEK
1F16..1F17  ; UNASSIGNED  # <reserved>..<reserved>
1F18..1F1D  ; DISALLOWED  # GREEK CAPITAL LETTER EPSILON WITH PSILI..GRE
1F1E..1F1F  ; UNASSIGNED  # <reserved>..<reserved>
1F20..1F27  ; PVALID      # GREEK SMALL LETTER ETA WITH PSILI..GREEK SMA
1F28..1F2F  ; DISALLOWED  # GREEK CAPITAL LETTER ETA WITH PSILI..GREEK C
1F30..1F37  ; PVALID      # GREEK SMALL LETTER IOTA WITH PSILI..GREEK SM
1F38..1F3F  ; DISALLOWED  # GREEK CAPITAL LETTER IOTA WITH PSILI..GREEK
1F40..1F45  ; PVALID      # GREEK SMALL LETTER OMICRON WITH PSILI..GREEK
1F46..1F47  ; UNASSIGNED  # <reserved>..<reserved>
1F48..1F4D  ; DISALLOWED  # GREEK CAPITAL LETTER OMICRON WITH PSILI..GRE
1F4E..1F4F  ; UNASSIGNED  # <reserved>..<reserved>
1F50..1F57  ; PVALID      # GREEK SMALL LETTER UPSILON WITH PSILI..GREEK
1F58        ; UNASSIGNED  # <reserved>
1F59        ; DISALLOWED  # GREEK CAPITAL LETTER UPSILON WITH DASIA
1F5A        ; UNASSIGNED  # <reserved>
1F5B        ; DISALLOWED  # GREEK CAPITAL LETTER UPSILON WITH DASIA AND
1F5C        ; UNASSIGNED  # <reserved>
1F5D        ; DISALLOWED  # GREEK CAPITAL LETTER UPSILON WITH DASIA AND
1F5E        ; UNASSIGNED  # <reserved>
1F5F        ; DISALLOWED  # GREEK CAPITAL LETTER UPSILON WITH DASIA AND
1F60..1F67  ; PVALID      # GREEK SMALL LETTER OMEGA WITH PSILI..GREEK S
1F68..1F6F  ; DISALLOWED  # GREEK CAPITAL LETTER OMEGA WITH PSILI..GREEK
1F70        ; PVALID      # GREEK SMALL LETTER ALPHA WITH VARIA
1F71        ; DISALLOWED  # GREEK SMALL LETTER ALPHA WITH OXIA
1F72        ; PVALID      # GREEK SMALL LETTER EPSILON WITH VARIA
1F73        ; DISALLOWED  # GREEK SMALL LETTER EPSILON WITH OXIA
1F74        ; PVALID      # GREEK SMALL LETTER ETA WITH VARIA
1F75        ; DISALLOWED  # GREEK SMALL LETTER ETA WITH OXIA
1F76        ; PVALID      # GREEK SMALL LETTER IOTA WITH VARIA
1F77        ; DISALLOWED  # GREEK SMALL LETTER IOTA WITH OXIA
1F78        ; PVALID      # GREEK SMALL LETTER OMICRON WITH VARIA
```

```
1F79          ; DISALLOWED  # GREEK SMALL LETTER OMICRON WITH OXIA
1F7A          ; PVALID      # GREEK SMALL LETTER UPSILON WITH VARIA
1F7B          ; DISALLOWED  # GREEK SMALL LETTER UPSILON WITH OXIA
1F7C          ; PVALID      # GREEK SMALL LETTER OMEGA WITH VARIA
1F7D          ; DISALLOWED  # GREEK SMALL LETTER OMEGA WITH OXIA
1F7E..1F7F    ; UNASSIGNED  # <reserved>..<reserved>
1F80..1FAF    ; DISALLOWED  # GREEK SMALL LETTER ALPHA WITH PSILI AND YPOG
1FB0..1FB1    ; PVALID      # GREEK SMALL LETTER ALPHA WITH VRACHY..GREEK
1FB2..1FB4    ; DISALLOWED  # GREEK SMALL LETTER ALPHA WITH VARIA AND YPOG
1FB5          ; UNASSIGNED  # <reserved>
1FB6          ; PVALID      # GREEK SMALL LETTER ALPHA WITH PERISPOMENI
1FB7..1FC4    ; DISALLOWED  # GREEK SMALL LETTER ALPHA WITH PERISPOMENI AN
1FC5          ; UNASSIGNED  # <reserved>
1FC6          ; PVALID      # GREEK SMALL LETTER ETA WITH PERISPOMENI
1FC7..1FCF    ; DISALLOWED  # GREEK SMALL LETTER ETA WITH PERISPOMENI AND
1FD0..1FD2    ; PVALID      # GREEK SMALL LETTER IOTA WITH VRACHY..GREEK S
1FD3          ; DISALLOWED  # GREEK SMALL LETTER IOTA WITH DIALYTIKA AND O
1FD4..1FD5    ; UNASSIGNED  # <reserved>..<reserved>
1FD6..1FD7    ; PVALID      # GREEK SMALL LETTER IOTA WITH PERISPOMENI..GR
1FD8..1FDB    ; DISALLOWED  # GREEK CAPITAL LETTER IOTA WITH VRACHY..GREEK
1FDC          ; UNASSIGNED  # <reserved>
1FDD..1FDF    ; DISALLOWED  # GREEK DASIA AND VARIA..GREEK DASIA AND PERIS
1FE0..1FE2    ; PVALID      # GREEK SMALL LETTER UPSILON WITH VRACHY..GREE
1FE3          ; DISALLOWED  # GREEK SMALL LETTER UPSILON WITH DIALYTIKA AN
1FE4..1FE7    ; PVALID      # GREEK SMALL LETTER RHO WITH PSILI..GREEK SMA
1FE8..1FEF    ; DISALLOWED  # GREEK CAPITAL LETTER UPSILON WITH VRACHY..GR
1FF0..1FF1    ; UNASSIGNED  # <reserved>..<reserved>
1FF2..1FF4    ; DISALLOWED  # GREEK SMALL LETTER OMEGA WITH VARIA AND YPOG
1FF5          ; UNASSIGNED  # <reserved>
1FF6          ; PVALID      # GREEK SMALL LETTER OMEGA WITH PERISPOMENI
1FF7..1FFE    ; DISALLOWED  # GREEK SMALL LETTER OMEGA WITH PERISPOMENI AN
1FFF          ; UNASSIGNED  # <reserved>
2000..200B    ; DISALLOWED  # EN QUAD..ZERO WIDTH SPACE
200C..200D    ; CONTEXTJ    # ZERO WIDTH NON-JOINER..ZERO WIDTH JOINER
200E..2064    ; DISALLOWED  # LEFT-TO-RIGHT MARK..INVISIBLE PLUS
2065          ; UNASSIGNED  # <reserved>
2066..2071    ; DISALLOWED  # LEFT-TO-RIGHT ISOLATE..SUPERSCRIPT LATIN SMA
2072..2073    ; UNASSIGNED  # <reserved>..<reserved>
2074..208E    ; DISALLOWED  # SUPERSCRIPT FOUR..SUBSCRIPT RIGHT PARENTHESI
208F          ; UNASSIGNED  # <reserved>
2090..209C    ; DISALLOWED  # LATIN SUBSCRIPT SMALL LETTER A..LATIN SUBSCR
209D..209F    ; UNASSIGNED  # <reserved>..<reserved>
20A0..20BF    ; DISALLOWED  # EURO-CURRENCY SIGN..BITCOIN SIGN
20C0..20CF    ; UNASSIGNED  # <reserved>..<reserved>
20D0..20F0    ; DISALLOWED  # COMBINING LEFT HARPOON ABOVE..COMBINING ASTE
20F1..20FF    ; UNASSIGNED  # <reserved>..<reserved>
2100..214D    ; DISALLOWED  # ACCOUNT OF..AKTIESELSKAB
214E          ; PVALID      # TURNED SMALL F
```

```
214F..2183  ; DISALLOWED  # SYMBOL FOR SAMARITAN SOURCE..ROMAN NUMERAL R
2184        ; PVALID      # LATIN SMALL LETTER REVERSED C
2185..218B  ; DISALLOWED  # ROMAN NUMERAL SIX LATE FORM..TURNED DIGIT TH
218C..218F  ; UNASSIGNED  # <reserved>..<reserved>
2190..2426  ; DISALLOWED  # LEFTWARDS ARROW..SYMBOL FOR SUBSTITUTE FORM
2427..243F  ; UNASSIGNED  # <reserved>..<reserved>
2440..244A  ; DISALLOWED  # OCR HOOK..OCR DOUBLE BACKSLASH
244B..245F  ; UNASSIGNED  # <reserved>..<reserved>
2460..2B73  ; DISALLOWED  # CIRCLED DIGIT ONE..DOWNWARDS TRIANGLE-HEADED
2B74..2B75  ; UNASSIGNED  # <reserved>..<reserved>
2B76..2B95  ; DISALLOWED  # NORTH WEST TRIANGLE-HEADED ARROW TO BAR..RIG
2B96..2B97  ; UNASSIGNED  # <reserved>..<reserved>
2B98..2BC8  ; DISALLOWED  # THREE-D TOP-LIGHTED LEFTWARDS EQUILATERAL AR
2BC9        ; UNASSIGNED  # <reserved>
2BCA..2BFE  ; DISALLOWED  # TOP HALF BLACK CIRCLE..REVERSED RIGHT ANGLE
2BFF        ; UNASSIGNED  # <reserved>
2C00..2C2E  ; DISALLOWED  # GLAGOLITIC CAPITAL LETTER AZU..GLAGOLITIC CA
2C2F        ; UNASSIGNED  # <reserved>
2C30..2C5E  ; PVALID      # GLAGOLITIC SMALL LETTER AZU..GLAGOLITIC SMAL
2C5F        ; UNASSIGNED  # <reserved>
2C60        ; DISALLOWED  # LATIN CAPITAL LETTER L WITH DOUBLE BAR
2C61        ; PVALID      # LATIN SMALL LETTER L WITH DOUBLE BAR
2C62..2C64  ; DISALLOWED  # LATIN CAPITAL LETTER L WITH MIDDLE TILDE..LA
2C65..2C66  ; PVALID      # LATIN SMALL LETTER A WITH STROKE..LATIN SMAL
2C67        ; DISALLOWED  # LATIN CAPITAL LETTER H WITH DESCENDER
2C68        ; PVALID      # LATIN SMALL LETTER H WITH DESCENDER
2C69        ; DISALLOWED  # LATIN CAPITAL LETTER K WITH DESCENDER
2C6A        ; PVALID      # LATIN SMALL LETTER K WITH DESCENDER
2C6B        ; DISALLOWED  # LATIN CAPITAL LETTER Z WITH DESCENDER
2C6C        ; PVALID      # LATIN SMALL LETTER Z WITH DESCENDER
2C6D..2C70  ; DISALLOWED  # LATIN CAPITAL LETTER ALPHA..LATIN CAPITAL LE
2C71        ; PVALID      # LATIN SMALL LETTER V WITH RIGHT HOOK
2C72        ; DISALLOWED  # LATIN CAPITAL LETTER W WITH HOOK
2C73..2C74  ; PVALID      # LATIN SMALL LETTER W WITH HOOK..LATIN SMALL
2C75        ; DISALLOWED  # LATIN CAPITAL LETTER HALF H
2C76..2C7B  ; PVALID      # LATIN SMALL LETTER HALF H..LATIN LETTER SMAL
2C7C..2C80  ; DISALLOWED  # LATIN SUBSCRIPT SMALL LETTER J..COPTIC CAPIT
2C81        ; PVALID      # COPTIC SMALL LETTER ALFA
2C82        ; DISALLOWED  # COPTIC CAPITAL LETTER VIDA
2C83        ; PVALID      # COPTIC SMALL LETTER VIDA
2C84        ; DISALLOWED  # COPTIC CAPITAL LETTER GAMMA
2C85        ; PVALID      # COPTIC SMALL LETTER GAMMA
2C86        ; DISALLOWED  # COPTIC CAPITAL LETTER DALDA
2C87        ; PVALID      # COPTIC SMALL LETTER DALDA
2C88        ; DISALLOWED  # COPTIC CAPITAL LETTER EIE
2C89        ; PVALID      # COPTIC SMALL LETTER EIE
2C8A        ; DISALLOWED  # COPTIC CAPITAL LETTER SOU
2C8B        ; PVALID      # COPTIC SMALL LETTER SOU
```

```
2C8C        ; DISALLOWED  # COPTIC CAPITAL LETTER ZATA
2C8D        ; PVALID      # COPTIC SMALL LETTER ZATA
2C8E        ; DISALLOWED  # COPTIC CAPITAL LETTER HATE
2C8F        ; PVALID      # COPTIC SMALL LETTER HATE
2C90        ; DISALLOWED  # COPTIC CAPITAL LETTER THETHE
2C91        ; PVALID      # COPTIC SMALL LETTER THETHE
2C92        ; DISALLOWED  # COPTIC CAPITAL LETTER IAUDA
2C93        ; PVALID      # COPTIC SMALL LETTER IAUDA
2C94        ; DISALLOWED  # COPTIC CAPITAL LETTER KAPA
2C95        ; PVALID      # COPTIC SMALL LETTER KAPA
2C96        ; DISALLOWED  # COPTIC CAPITAL LETTER LAULA
2C97        ; PVALID      # COPTIC SMALL LETTER LAULA
2C98        ; DISALLOWED  # COPTIC CAPITAL LETTER MI
2C99        ; PVALID      # COPTIC SMALL LETTER MI
2C9A        ; DISALLOWED  # COPTIC CAPITAL LETTER NI
2C9B        ; PVALID      # COPTIC SMALL LETTER NI
2C9C        ; DISALLOWED  # COPTIC CAPITAL LETTER KSI
2C9D        ; PVALID      # COPTIC SMALL LETTER KSI
2C9E        ; DISALLOWED  # COPTIC CAPITAL LETTER O
2C9F        ; PVALID      # COPTIC SMALL LETTER O
2CA0        ; DISALLOWED  # COPTIC CAPITAL LETTER PI
2CA1        ; PVALID      # COPTIC SMALL LETTER PI
2CA2        ; DISALLOWED  # COPTIC CAPITAL LETTER RO
2CA3        ; PVALID      # COPTIC SMALL LETTER RO
2CA4        ; DISALLOWED  # COPTIC CAPITAL LETTER SIMA
2CA5        ; PVALID      # COPTIC SMALL LETTER SIMA
2CA6        ; DISALLOWED  # COPTIC CAPITAL LETTER TAU
2CA7        ; PVALID      # COPTIC SMALL LETTER TAU
2CA8        ; DISALLOWED  # COPTIC CAPITAL LETTER UA
2CA9        ; PVALID      # COPTIC SMALL LETTER UA
2CAA        ; DISALLOWED  # COPTIC CAPITAL LETTER FI
2CAB        ; PVALID      # COPTIC SMALL LETTER FI
2CAC        ; DISALLOWED  # COPTIC CAPITAL LETTER KHI
2CAD        ; PVALID      # COPTIC SMALL LETTER KHI
2CAE        ; DISALLOWED  # COPTIC CAPITAL LETTER PSI
2CAF        ; PVALID      # COPTIC SMALL LETTER PSI
2CB0        ; DISALLOWED  # COPTIC CAPITAL LETTER OOU
2CB1        ; PVALID      # COPTIC SMALL LETTER OOU
2CB2        ; DISALLOWED  # COPTIC CAPITAL LETTER DIALECT-P ALEF
2CB3        ; PVALID      # COPTIC SMALL LETTER DIALECT-P ALEF
2CB4        ; DISALLOWED  # COPTIC CAPITAL LETTER OLD COPTIC AIN
2CB5        ; PVALID      # COPTIC SMALL LETTER OLD COPTIC AIN
2CB6        ; DISALLOWED  # COPTIC CAPITAL LETTER CRYPTOGRAMMIC EIE
2CB7        ; PVALID      # COPTIC SMALL LETTER CRYPTOGRAMMIC EIE
2CB8        ; DISALLOWED  # COPTIC CAPITAL LETTER DIALECT-P KAPA
2CB9        ; PVALID      # COPTIC SMALL LETTER DIALECT-P KAPA
2CBA        ; DISALLOWED  # COPTIC CAPITAL LETTER DIALECT-P NI
2CBB        ; PVALID      # COPTIC SMALL LETTER DIALECT-P NI
```

```
2CBC        ; DISALLOWED  # COPTIC CAPITAL LETTER CRYPTOGRAMMIC NI
2CBD        ; PVALID      # COPTIC SMALL LETTER CRYPTOGRAMMIC NI
2CBE        ; DISALLOWED  # COPTIC CAPITAL LETTER OLD COPTIC OOU
2CBF        ; PVALID      # COPTIC SMALL LETTER OLD COPTIC OOU
2CC0        ; DISALLOWED  # COPTIC CAPITAL LETTER SAMPI
2CC1        ; PVALID      # COPTIC SMALL LETTER SAMPI
2CC2        ; DISALLOWED  # COPTIC CAPITAL LETTER CROSSED SHEI
2CC3        ; PVALID      # COPTIC SMALL LETTER CROSSED SHEI
2CC4        ; DISALLOWED  # COPTIC CAPITAL LETTER OLD COPTIC SHEI
2CC5        ; PVALID      # COPTIC SMALL LETTER OLD COPTIC SHEI
2CC6        ; DISALLOWED  # COPTIC CAPITAL LETTER OLD COPTIC ESH
2CC7        ; PVALID      # COPTIC SMALL LETTER OLD COPTIC ESH
2CC8        ; DISALLOWED  # COPTIC CAPITAL LETTER AKHMIMIC KHEI
2CC9        ; PVALID      # COPTIC SMALL LETTER AKHMIMIC KHEI
2CCA        ; DISALLOWED  # COPTIC CAPITAL LETTER DIALECT-P HORI
2CCB        ; PVALID      # COPTIC SMALL LETTER DIALECT-P HORI
2CCC        ; DISALLOWED  # COPTIC CAPITAL LETTER OLD COPTIC HORI
2CCD        ; PVALID      # COPTIC SMALL LETTER OLD COPTIC HORI
2CCE        ; DISALLOWED  # COPTIC CAPITAL LETTER OLD COPTIC HA
2CCF        ; PVALID      # COPTIC SMALL LETTER OLD COPTIC HA
2CD0        ; DISALLOWED  # COPTIC CAPITAL LETTER L-SHAPED HA
2CD1        ; PVALID      # COPTIC SMALL LETTER L-SHAPED HA
2CD2        ; DISALLOWED  # COPTIC CAPITAL LETTER OLD COPTIC HEI
2CD3        ; PVALID      # COPTIC SMALL LETTER OLD COPTIC HEI
2CD4        ; DISALLOWED  # COPTIC CAPITAL LETTER OLD COPTIC HAT
2CD5        ; PVALID      # COPTIC SMALL LETTER OLD COPTIC HAT
2CD6        ; DISALLOWED  # COPTIC CAPITAL LETTER OLD COPTIC GANGIA
2CD7        ; PVALID      # COPTIC SMALL LETTER OLD COPTIC GANGIA
2CD8        ; DISALLOWED  # COPTIC CAPITAL LETTER OLD COPTIC DJA
2CD9        ; PVALID      # COPTIC SMALL LETTER OLD COPTIC DJA
2CDA        ; DISALLOWED  # COPTIC CAPITAL LETTER OLD COPTIC SHIMA
2CDB        ; PVALID      # COPTIC SMALL LETTER OLD COPTIC SHIMA
2CDC        ; DISALLOWED  # COPTIC CAPITAL LETTER OLD NUBIAN SHIMA
2CDD        ; PVALID      # COPTIC SMALL LETTER OLD NUBIAN SHIMA
2CDE        ; DISALLOWED  # COPTIC CAPITAL LETTER OLD NUBIAN NGI
2CDF        ; PVALID      # COPTIC SMALL LETTER OLD NUBIAN NGI
2CE0        ; DISALLOWED  # COPTIC CAPITAL LETTER OLD NUBIAN NYI
2CE1        ; PVALID      # COPTIC SMALL LETTER OLD NUBIAN NYI
2CE2        ; DISALLOWED  # COPTIC CAPITAL LETTER OLD NUBIAN WAU
2CE3..2CE4  ; PVALID      # COPTIC SMALL LETTER OLD NUBIAN WAU..COPTIC S
2CE5..2CEB  ; DISALLOWED  # COPTIC SYMBOL MI RO..COPTIC CAPITAL LETTER C
2CEC        ; PVALID      # COPTIC SMALL LETTER CRYPTOGRAMMIC SHEI
2CED        ; DISALLOWED  # COPTIC CAPITAL LETTER CRYPTOGRAMMIC GANGIA
2CEE..2CF1  ; PVALID      # COPTIC SMALL LETTER CRYPTOGRAMMIC GANGIA..CO
2CF2        ; DISALLOWED  # COPTIC CAPITAL LETTER BOHAIRIC KHEI
2CF3        ; PVALID      # COPTIC SMALL LETTER BOHAIRIC KHEI
2CF4..2CF8  ; UNASSIGNED  # <reserved>..<reserved>
2CF9..2CFF  ; DISALLOWED  # COPTIC OLD NUBIAN FULL STOP..COPTIC MORPHOLO
```

```
2D00..2D25  ; PVALID      # GEORGIAN SMALL LETTER AN..GEORGIAN SMALL LET
2D26        ; UNASSIGNED  # <reserved>
2D27        ; PVALID      # GEORGIAN SMALL LETTER YN
2D28..2D2C  ; UNASSIGNED  # <reserved>..<reserved>
2D2D        ; PVALID      # GEORGIAN SMALL LETTER AEN
2D2E..2D2F  ; UNASSIGNED  # <reserved>..<reserved>
2D30..2D67  ; PVALID      # TIFINAGH LETTER YA..TIFINAGH LETTER YO
2D68..2D6E  ; UNASSIGNED  # <reserved>..<reserved>
2D6F..2D70  ; DISALLOWED  # TIFINAGH MODIFIER LETTER LABIALIZATION MARK.
2D71..2D7E  ; UNASSIGNED  # <reserved>..<reserved>
2D7F..2D96  ; PVALID      # TIFINAGH CONSONANT JOINER..ETHIOPIC SYLLABLE
2D97..2D9F  ; UNASSIGNED  # <reserved>..<reserved>
2DA0..2DA6  ; PVALID      # ETHIOPIC SYLLABLE SSA..ETHIOPIC SYLLABLE SSO
2DA7        ; UNASSIGNED  # <reserved>
2DA8..2DAE  ; PVALID      # ETHIOPIC SYLLABLE CCA..ETHIOPIC SYLLABLE CCO
2DAF        ; UNASSIGNED  # <reserved>
2DB0..2DB6  ; PVALID      # ETHIOPIC SYLLABLE ZZA..ETHIOPIC SYLLABLE ZZO
2DB7        ; UNASSIGNED  # <reserved>
2DB8..2DBE  ; PVALID      # ETHIOPIC SYLLABLE CCHA..ETHIOPIC SYLLABLE CC
2DBF        ; UNASSIGNED  # <reserved>
2DC0..2DC6  ; PVALID      # ETHIOPIC SYLLABLE QYA..ETHIOPIC SYLLABLE QYO
2DC7        ; UNASSIGNED  # <reserved>
2DC8..2DCE  ; PVALID      # ETHIOPIC SYLLABLE KYA..ETHIOPIC SYLLABLE KYO
2DCF        ; UNASSIGNED  # <reserved>
2DD0..2DD6  ; PVALID      # ETHIOPIC SYLLABLE XYA..ETHIOPIC SYLLABLE XYO
2DD7        ; UNASSIGNED  # <reserved>
2DD8..2DDE  ; PVALID      # ETHIOPIC SYLLABLE GYA..ETHIOPIC SYLLABLE GYO
2DDF        ; UNASSIGNED  # <reserved>
2DE0..2DFF  ; PVALID      # COMBINING CYRILLIC LETTER BE..COMBINING CYRI
2E00..2E2E  ; DISALLOWED  # RIGHT ANGLE SUBSTITUTION MARKER..REVERSED QU
2E2F        ; PVALID      # VERTICAL TILDE
2E30..2E4E  ; DISALLOWED  # RING POINT..PUNCTUS ELEVATUS MARK
2E4F..2E7F  ; UNASSIGNED  # <reserved>..<reserved>
2E80..2E99  ; DISALLOWED  # CJK RADICAL REPEAT..CJK RADICAL RAP
2E9A        ; UNASSIGNED  # <reserved>
2E9B..2EF3  ; DISALLOWED  # CJK RADICAL CHOKE..CJK RADICAL C-SIMPLIFIED
2EF4..2EFF  ; UNASSIGNED  # <reserved>..<reserved>
2F00..2FD5  ; DISALLOWED  # KANGXI RADICAL ONE..KANGXI RADICAL FLUTE
2FD6..2FEF  ; UNASSIGNED  # <reserved>..<reserved>
2FF0..2FFB  ; DISALLOWED  # IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO RI
2FFC..2FFF  ; UNASSIGNED  # <reserved>..<reserved>
3000..3004  ; DISALLOWED  # IDEOGRAPHIC SPACE..JAPANESE INDUSTRIAL STAND
3005..3007  ; PVALID      # IDEOGRAPHIC ITERATION MARK..IDEOGRAPHIC NUMB
3008..3029  ; DISALLOWED  # LEFT ANGLE BRACKET..HANGZHOU NUMERAL NINE
302A..302D  ; PVALID      # IDEOGRAPHIC LEVEL TONE MARK..IDEOGRAPHIC ENT
302E..303B  ; DISALLOWED  # HANGUL SINGLE DOT TONE MARK..VERTICAL IDEOGR
303C        ; PVALID      # MASU MARK
303D..303F  ; DISALLOWED  # PART ALTERNATION MARK..IDEOGRAPHIC HALF FILL
```

```
3040        ; UNASSIGNED  # <reserved>
3041..3096  ; PVALID      # HIRAGANA LETTER SMALL A..HIRAGANA LETTER SMA
3097..3098  ; UNASSIGNED  # <reserved>..<reserved>
3099..309A  ; PVALID      # COMBINING KATAKANA-HIRAGANA VOICED SOUND MAR
309B..309C  ; DISALLOWED  # KATAKANA-HIRAGANA VOICED SOUND MARK..KATAKAN
309D..309E  ; PVALID      # HIRAGANA ITERATION MARK..HIRAGANA VOICED ITE
309F..30A0  ; DISALLOWED  # HIRAGANA DIGRAPH YORI..KATAKANA-HIRAGANA DOU
30A1..30FA  ; PVALID      # KATAKANA LETTER SMALL A..KATAKANA LETTER VO
30FB        ; CONTEXTO    # KATAKANA MIDDLE DOT
30FC..30FE  ; PVALID      # KATAKANA-HIRAGANA PROLONGED SOUND MARK..KATA
30FF        ; DISALLOWED  # KATAKANA DIGRAPH KOTO
3100..3104  ; UNASSIGNED  # <reserved>..<reserved>
3105..312F  ; PVALID      # BOPOMOFO LETTER B..BOPOMOFO LETTER NN
3130        ; UNASSIGNED  # <reserved>
3131..318E  ; DISALLOWED  # HANGUL LETTER KIYEOK..HANGUL LETTER ARAEAE
318F        ; UNASSIGNED  # <reserved>
3190..319F  ; DISALLOWED  # IDEOGRAPHIC ANNOTATION LINKING MARK..IDEOGRA
31A0..31BA  ; PVALID      # BOPOMOFO LETTER BU..BOPOMOFO LETTER ZY
31BB..31BF  ; UNASSIGNED  # <reserved>..<reserved>
31C0..31E3  ; DISALLOWED  # CJK STROKE T..CJK STROKE Q
31E4..31EF  ; UNASSIGNED  # <reserved>..<reserved>
31F0..31FF  ; PVALID      # KATAKANA LETTER SMALL KU..KATAKANA LETTER SM
3200..321E  ; DISALLOWED  # PARENTHESIZED HANGUL KIYEOK..PARENTHESIZED K
321F        ; UNASSIGNED  # <reserved>
3220..32FE  ; DISALLOWED  # PARENTHESIZED IDEOGRAPH ONE..CIRCLED KATAKAN
32FF        ; UNASSIGNED  # <reserved>
3300..33FF  ; DISALLOWED  # SQUARE APAATO..SQUARE GAL
3400..4DB5  ; PVALID      # <CJK Ideograph Extension A>..<CJK Ideograph
4DB6..4DBF  ; UNASSIGNED  # <reserved>..<reserved>
4DC0..4DFF  ; DISALLOWED  # HEXAGRAM FOR THE CREATIVE HEAVEN..HEXAGRAM F
4E00..9FEF  ; PVALID      # <CJK Ideograph>..<CJK Ideograph>
9FF0..9FFF  ; UNASSIGNED  # <reserved>..<reserved>
A000..A48C  ; PVALID      # YI SYLLABLE IT..YI SYLLABLE YYR
A48D..A48F  ; UNASSIGNED  # <reserved>..<reserved>
A490..A4C6  ; DISALLOWED  # YI RADICAL QOT..YI RADICAL KE
A4C7..A4CF  ; UNASSIGNED  # <reserved>..<reserved>
A4D0..A4FD  ; PVALID      # LISU LETTER BA..LISU LETTER TONE MYA JEU
A4FE..A4FF  ; DISALLOWED  # LISU PUNCTUATION COMMA..LISU PUNCTUATION FUL
A500..A60C  ; PVALID      # VAI SYLLABLE EE..VAI SYLLABLE LENGTHENER
A60D..A60F  ; DISALLOWED  # VAI COMMA..VAI QUESTION MARK
A610..A62B  ; PVALID      # VAI SYLLABLE NDOLE FA..VAI SYLLABLE NDOLE DO
A62C..A63F  ; UNASSIGNED  # <reserved>..<reserved>
A640        ; DISALLOWED  # CYRILLIC CAPITAL LETTER ZEMLYA
A641        ; PVALID      # CYRILLIC SMALL LETTER ZEMLYA
A642        ; DISALLOWED  # CYRILLIC CAPITAL LETTER DZELO
A643        ; PVALID      # CYRILLIC SMALL LETTER DZELO
A644        ; DISALLOWED  # CYRILLIC CAPITAL LETTER REVERSED DZE
A645        ; PVALID      # CYRILLIC SMALL LETTER REVERSED DZE
```

```
A646          ; DISALLOWED  # CYRILLIC CAPITAL LETTER IOTA
A647          ; PVALID      # CYRILLIC SMALL LETTER IOTA
A648          ; DISALLOWED  # CYRILLIC CAPITAL LETTER DJERV
A649          ; PVALID      # CYRILLIC SMALL LETTER DJERV
A64A          ; DISALLOWED  # CYRILLIC CAPITAL LETTER MONOGRAPH UK
A64B          ; PVALID      # CYRILLIC SMALL LETTER MONOGRAPH UK
A64C          ; DISALLOWED  # CYRILLIC CAPITAL LETTER BROAD OMEGA
A64D          ; PVALID      # CYRILLIC SMALL LETTER BROAD OMEGA
A64E          ; DISALLOWED  # CYRILLIC CAPITAL LETTER NEUTRAL YER
A64F          ; PVALID      # CYRILLIC SMALL LETTER NEUTRAL YER
A650          ; DISALLOWED  # CYRILLIC CAPITAL LETTER YERU WITH BACK YER
A651          ; PVALID      # CYRILLIC SMALL LETTER YERU WITH BACK YER
A652          ; DISALLOWED  # CYRILLIC CAPITAL LETTER IOTIFIED YAT
A653          ; PVALID      # CYRILLIC SMALL LETTER IOTIFIED YAT
A654          ; DISALLOWED  # CYRILLIC CAPITAL LETTER REVERSED YU
A655          ; PVALID      # CYRILLIC SMALL LETTER REVERSED YU
A656          ; DISALLOWED  # CYRILLIC CAPITAL LETTER IOTIFIED A
A657          ; PVALID      # CYRILLIC SMALL LETTER IOTIFIED A
A658          ; DISALLOWED  # CYRILLIC CAPITAL LETTER CLOSED LITTLE YUS
A659          ; PVALID      # CYRILLIC SMALL LETTER CLOSED LITTLE YUS
A65A          ; DISALLOWED  # CYRILLIC CAPITAL LETTER BLENDED YUS
A65B          ; PVALID      # CYRILLIC SMALL LETTER BLENDED YUS
A65C          ; DISALLOWED  # CYRILLIC CAPITAL LETTER IOTIFIED CLOSED LITT
A65D          ; PVALID      # CYRILLIC SMALL LETTER IOTIFIED CLOSED LITTLE
A65E          ; DISALLOWED  # CYRILLIC CAPITAL LETTER YN
A65F          ; PVALID      # CYRILLIC SMALL LETTER YN
A660          ; DISALLOWED  # CYRILLIC CAPITAL LETTER REVERSED TSE
A661          ; PVALID      # CYRILLIC SMALL LETTER REVERSED TSE
A662          ; DISALLOWED  # CYRILLIC CAPITAL LETTER SOFT DE
A663          ; PVALID      # CYRILLIC SMALL LETTER SOFT DE
A664          ; DISALLOWED  # CYRILLIC CAPITAL LETTER SOFT EL
A665          ; PVALID      # CYRILLIC SMALL LETTER SOFT EL
A666          ; DISALLOWED  # CYRILLIC CAPITAL LETTER SOFT EM
A667          ; PVALID      # CYRILLIC SMALL LETTER SOFT EM
A668          ; DISALLOWED  # CYRILLIC CAPITAL LETTER MONOCULAR O
A669          ; PVALID      # CYRILLIC SMALL LETTER MONOCULAR O
A66A          ; DISALLOWED  # CYRILLIC CAPITAL LETTER BINOCULAR O
A66B          ; PVALID      # CYRILLIC SMALL LETTER BINOCULAR O
A66C          ; DISALLOWED  # CYRILLIC CAPITAL LETTER DOUBLE MONOCULAR O
A66D..A66F    ; PVALID      # CYRILLIC SMALL LETTER DOUBLE MONOCULAR O..CO
A670..A673    ; DISALLOWED  # COMBINING CYRILLIC TEN MILLIONS SIGN..SLAVON
A674..A67D    ; PVALID      # COMBINING CYRILLIC LETTER UKRAINIAN IE..COMB
A67E          ; DISALLOWED  # CYRILLIC KAVYKA
A67F          ; PVALID      # CYRILLIC PAYEROK
A680          ; DISALLOWED  # CYRILLIC CAPITAL LETTER DWE
A681          ; PVALID      # CYRILLIC SMALL LETTER DWE
A682          ; DISALLOWED  # CYRILLIC CAPITAL LETTER DZWE
A683          ; PVALID      # CYRILLIC SMALL LETTER DZWE
```

```
A684          ; DISALLOWED  # CYRILLIC CAPITAL LETTER ZHWE
A685          ; PVALID      # CYRILLIC SMALL LETTER ZHWE
A686          ; DISALLOWED  # CYRILLIC CAPITAL LETTER CCHE
A687          ; PVALID      # CYRILLIC SMALL LETTER CCHE
A688          ; DISALLOWED  # CYRILLIC CAPITAL LETTER DZZE
A689          ; PVALID      # CYRILLIC SMALL LETTER DZZE
A68A          ; DISALLOWED  # CYRILLIC CAPITAL LETTER TE WITH MIDDLE HOOK
A68B          ; PVALID      # CYRILLIC SMALL LETTER TE WITH MIDDLE HOOK
A68C          ; DISALLOWED  # CYRILLIC CAPITAL LETTER TWE
A68D          ; PVALID      # CYRILLIC SMALL LETTER TWE
A68E          ; DISALLOWED  # CYRILLIC CAPITAL LETTER TSWE
A68F          ; PVALID      # CYRILLIC SMALL LETTER TSWE
A690          ; DISALLOWED  # CYRILLIC CAPITAL LETTER TSSE
A691          ; PVALID      # CYRILLIC SMALL LETTER TSSE
A692          ; DISALLOWED  # CYRILLIC CAPITAL LETTER TCHE
A693          ; PVALID      # CYRILLIC SMALL LETTER TCHE
A694          ; DISALLOWED  # CYRILLIC CAPITAL LETTER HWE
A695          ; PVALID      # CYRILLIC SMALL LETTER HWE
A696          ; DISALLOWED  # CYRILLIC CAPITAL LETTER SHWE
A697          ; PVALID      # CYRILLIC SMALL LETTER SHWE
A698          ; DISALLOWED  # CYRILLIC CAPITAL LETTER DOUBLE O
A699          ; PVALID      # CYRILLIC SMALL LETTER DOUBLE O
A69A          ; DISALLOWED  # CYRILLIC CAPITAL LETTER CROSSED O
A69B          ; PVALID      # CYRILLIC SMALL LETTER CROSSED O
A69C..A69D    ; DISALLOWED  # MODIFIER LETTER CYRILLIC HARD SIGN..MODIFIER
A69E..A6E5    ; PVALID      # COMBINING CYRILLIC LETTER EF..BAMUM LETTER K
A6E6..A6EF    ; DISALLOWED  # BAMUM LETTER MO..BAMUM LETTER KOGHOM
A6F0..A6F1    ; PVALID      # BAMUM COMBINING MARK KOQNDON..BAMUM COMBININ
A6F2..A6F7    ; DISALLOWED  # BAMUM NJAEMLI..BAMUM QUESTION MARK
A6F8..A6FF    ; UNASSIGNED  # <reserved>..<reserved>
A700..A716    ; DISALLOWED  # MODIFIER LETTER CHINESE TONE YIN PING..MODIF
A717..A71F    ; PVALID      # MODIFIER LETTER DOT VERTICAL BAR..MODIFIER L
A720..A722    ; DISALLOWED  # MODIFIER LETTER STRESS AND HIGH TONE..LATIN
A723          ; PVALID      # LATIN SMALL LETTER EGYPTOLOGICAL ALEF
A724          ; DISALLOWED  # LATIN CAPITAL LETTER EGYPTOLOGICAL AIN
A725          ; PVALID      # LATIN SMALL LETTER EGYPTOLOGICAL AIN
A726          ; DISALLOWED  # LATIN CAPITAL LETTER HENG
A727          ; PVALID      # LATIN SMALL LETTER HENG
A728          ; DISALLOWED  # LATIN CAPITAL LETTER TZ
A729          ; PVALID      # LATIN SMALL LETTER TZ
A72A          ; DISALLOWED  # LATIN CAPITAL LETTER TRESILLO
A72B          ; PVALID      # LATIN SMALL LETTER TRESILLO
A72C          ; DISALLOWED  # LATIN CAPITAL LETTER CUATRILLO
A72D          ; PVALID      # LATIN SMALL LETTER CUATRILLO
A72E          ; DISALLOWED  # LATIN CAPITAL LETTER CUATRILLO WITH COMMA
A72F..A731    ; PVALID      # LATIN SMALL LETTER CUATRILLO WITH COMMA..LAT
A732          ; DISALLOWED  # LATIN CAPITAL LETTER AA
A733          ; PVALID      # LATIN SMALL LETTER AA
```

```
A734        ; DISALLOWED  # LATIN CAPITAL LETTER AO
A735        ; PVALID      # LATIN SMALL LETTER AO
A736        ; DISALLOWED  # LATIN CAPITAL LETTER AU
A737        ; PVALID      # LATIN SMALL LETTER AU
A738        ; DISALLOWED  # LATIN CAPITAL LETTER AV
A739        ; PVALID      # LATIN SMALL LETTER AV
A73A        ; DISALLOWED  # LATIN CAPITAL LETTER AV WITH HORIZONTAL BAR
A73B        ; PVALID      # LATIN SMALL LETTER AV WITH HORIZONTAL BAR
A73C        ; DISALLOWED  # LATIN CAPITAL LETTER AY
A73D        ; PVALID      # LATIN SMALL LETTER AY
A73E        ; DISALLOWED  # LATIN CAPITAL LETTER REVERSED C WITH DOT
A73F        ; PVALID      # LATIN SMALL LETTER REVERSED C WITH DOT
A740        ; DISALLOWED  # LATIN CAPITAL LETTER K WITH STROKE
A741        ; PVALID      # LATIN SMALL LETTER K WITH STROKE
A742        ; DISALLOWED  # LATIN CAPITAL LETTER K WITH DIAGONAL STROKE
A743        ; PVALID      # LATIN SMALL LETTER K WITH DIAGONAL STROKE
A744        ; DISALLOWED  # LATIN CAPITAL LETTER K WITH STROKE AND DIAGO
A745        ; PVALID      # LATIN SMALL LETTER K WITH STROKE AND DIAGONA
A746        ; DISALLOWED  # LATIN CAPITAL LETTER BROKEN L
A747        ; PVALID      # LATIN SMALL LETTER BROKEN L
A748        ; DISALLOWED  # LATIN CAPITAL LETTER L WITH HIGH STROKE
A749        ; PVALID      # LATIN SMALL LETTER L WITH HIGH STROKE
A74A        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH LONG STROKE OVER
A74B        ; PVALID      # LATIN SMALL LETTER O WITH LONG STROKE OVERLA
A74C        ; DISALLOWED  # LATIN CAPITAL LETTER O WITH LOOP
A74D        ; PVALID      # LATIN SMALL LETTER O WITH LOOP
A74E        ; DISALLOWED  # LATIN CAPITAL LETTER OO
A74F        ; PVALID      # LATIN SMALL LETTER OO
A750        ; DISALLOWED  # LATIN CAPITAL LETTER P WITH STROKE THROUGH D
A751        ; PVALID      # LATIN SMALL LETTER P WITH STROKE THROUGH DES
A752        ; DISALLOWED  # LATIN CAPITAL LETTER P WITH FLOURISH
A753        ; PVALID      # LATIN SMALL LETTER P WITH FLOURISH
A754        ; DISALLOWED  # LATIN CAPITAL LETTER P WITH SQUIRREL TAIL
A755        ; PVALID      # LATIN SMALL LETTER P WITH SQUIRREL TAIL
A756        ; DISALLOWED  # LATIN CAPITAL LETTER Q WITH STROKE THROUGH D
A757        ; PVALID      # LATIN SMALL LETTER Q WITH STROKE THROUGH DES
A758        ; DISALLOWED  # LATIN CAPITAL LETTER Q WITH DIAGONAL STROKE
A759        ; PVALID      # LATIN SMALL LETTER Q WITH DIAGONAL STROKE
A75A        ; DISALLOWED  # LATIN CAPITAL LETTER R ROTUNDA
A75B        ; PVALID      # LATIN SMALL LETTER R ROTUNDA
A75C        ; DISALLOWED  # LATIN CAPITAL LETTER RUM ROTUNDA
A75D        ; PVALID      # LATIN SMALL LETTER RUM ROTUNDA
A75E        ; DISALLOWED  # LATIN CAPITAL LETTER V WITH DIAGONAL STROKE
A75F        ; PVALID      # LATIN SMALL LETTER V WITH DIAGONAL STROKE
A760        ; DISALLOWED  # LATIN CAPITAL LETTER VY
A761        ; PVALID      # LATIN SMALL LETTER VY
A762        ; DISALLOWED  # LATIN CAPITAL LETTER VISIGOTHIC Z
A763        ; PVALID      # LATIN SMALL LETTER VISIGOTHIC Z
```

```
A764          ; DISALLOWED  # LATIN CAPITAL LETTER THORN WITH STROKE
A765          ; PVALID      # LATIN SMALL LETTER THORN WITH STROKE
A766          ; DISALLOWED  # LATIN CAPITAL LETTER THORN WITH STROKE THROU
A767          ; PVALID      # LATIN SMALL LETTER THORN WITH STROKE THROUGH
A768          ; DISALLOWED  # LATIN CAPITAL LETTER VEND
A769          ; PVALID      # LATIN SMALL LETTER VEND
A76A          ; DISALLOWED  # LATIN CAPITAL LETTER ET
A76B          ; PVALID      # LATIN SMALL LETTER ET
A76C          ; DISALLOWED  # LATIN CAPITAL LETTER IS
A76D          ; PVALID      # LATIN SMALL LETTER IS
A76E          ; DISALLOWED  # LATIN CAPITAL LETTER CON
A76F          ; PVALID      # LATIN SMALL LETTER CON
A770          ; DISALLOWED  # MODIFIER LETTER US
A771..A778    ; PVALID      # LATIN SMALL LETTER DUM..LATIN SMALL LETTER U
A779          ; DISALLOWED  # LATIN CAPITAL LETTER INSULAR D
A77A          ; PVALID      # LATIN SMALL LETTER INSULAR D
A77B          ; DISALLOWED  # LATIN CAPITAL LETTER INSULAR F
A77C          ; PVALID      # LATIN SMALL LETTER INSULAR F
A77D..A77E    ; DISALLOWED  # LATIN CAPITAL LETTER INSULAR G..LATIN CAPITA
A77F          ; PVALID      # LATIN SMALL LETTER TURNED INSULAR G
A780          ; DISALLOWED  # LATIN CAPITAL LETTER TURNED L
A781          ; PVALID      # LATIN SMALL LETTER TURNED L
A782          ; DISALLOWED  # LATIN CAPITAL LETTER INSULAR R
A783          ; PVALID      # LATIN SMALL LETTER INSULAR R
A784          ; DISALLOWED  # LATIN CAPITAL LETTER INSULAR S
A785          ; PVALID      # LATIN SMALL LETTER INSULAR S
A786          ; DISALLOWED  # LATIN CAPITAL LETTER INSULAR T
A787..A788    ; PVALID      # LATIN SMALL LETTER INSULAR T..MODIFIER LETTE
A789..A78B    ; DISALLOWED  # MODIFIER LETTER COLON..LATIN CAPITAL LETTER
A78C          ; PVALID      # LATIN SMALL LETTER SALTILLO
A78D          ; DISALLOWED  # LATIN CAPITAL LETTER TURNED H
A78E..A78F    ; PVALID      # LATIN SMALL LETTER L WITH RETROFLEX HOOK AND
A790          ; DISALLOWED  # LATIN CAPITAL LETTER N WITH DESCENDER
A791          ; PVALID      # LATIN SMALL LETTER N WITH DESCENDER
A792          ; DISALLOWED  # LATIN CAPITAL LETTER C WITH BAR
A793..A795    ; PVALID      # LATIN SMALL LETTER C WITH BAR..LATIN SMALL L
A796          ; DISALLOWED  # LATIN CAPITAL LETTER B WITH FLOURISH
A797          ; PVALID      # LATIN SMALL LETTER B WITH FLOURISH
A798          ; DISALLOWED  # LATIN CAPITAL LETTER F WITH STROKE
A799          ; PVALID      # LATIN SMALL LETTER F WITH STROKE
A79A          ; DISALLOWED  # LATIN CAPITAL LETTER VOLAPUK AE
A79B          ; PVALID      # LATIN SMALL LETTER VOLAPUK AE
A79C          ; DISALLOWED  # LATIN CAPITAL LETTER VOLAPUK OE
A79D          ; PVALID      # LATIN SMALL LETTER VOLAPUK OE
A79E          ; DISALLOWED  # LATIN CAPITAL LETTER VOLAPUK UE
A79F          ; PVALID      # LATIN SMALL LETTER VOLAPUK UE
A7A0          ; DISALLOWED  # LATIN CAPITAL LETTER G WITH OBLIQUE STROKE
A7A1          ; PVALID      # LATIN SMALL LETTER G WITH OBLIQUE STROKE
```

```
A7A2        ; DISALLOWED  # LATIN CAPITAL LETTER K WITH OBLIQUE STROKE
A7A3        ; PVALID      # LATIN SMALL LETTER K WITH OBLIQUE STROKE
A7A4        ; DISALLOWED  # LATIN CAPITAL LETTER N WITH OBLIQUE STROKE
A7A5        ; PVALID      # LATIN SMALL LETTER N WITH OBLIQUE STROKE
A7A6        ; DISALLOWED  # LATIN CAPITAL LETTER R WITH OBLIQUE STROKE
A7A7        ; PVALID      # LATIN SMALL LETTER R WITH OBLIQUE STROKE
A7A8        ; DISALLOWED  # LATIN CAPITAL LETTER S WITH OBLIQUE STROKE
A7A9        ; PVALID      # LATIN SMALL LETTER S WITH OBLIQUE STROKE
A7AA..A7AE  ; DISALLOWED  # LATIN CAPITAL LETTER H WITH HOOK..LATIN CAPI
A7AF        ; PVALID      # LATIN LETTER SMALL CAPITAL Q
A7B0..A7B4  ; DISALLOWED  # LATIN CAPITAL LETTER TURNED K..LATIN CAPITAL
A7B5        ; PVALID      # LATIN SMALL LETTER BETA
A7B6        ; DISALLOWED  # LATIN CAPITAL LETTER OMEGA
A7B7        ; PVALID      # LATIN SMALL LETTER OMEGA
A7B8        ; DISALLOWED  # LATIN CAPITAL LETTER U WITH STROKE
A7B9        ; PVALID      # LATIN SMALL LETTER U WITH STROKE
A7BA..A7F6  ; UNASSIGNED  # <reserved>..<reserved>
A7F7        ; PVALID      # LATIN EPIGRAPHIC LETTER SIDEWAYS I
A7F8..A7F9  ; DISALLOWED  # MODIFIER LETTER CAPITAL H WITH STROKE..MODIF
A7FA..A827  ; PVALID      # LATIN LETTER SMALL CAPITAL TURNED M..SYLOTI
A828..A82B  ; DISALLOWED  # SYLOTI NAGRI POETRY MARK-1..SYLOTI NAGRI POE
A82C..A82F  ; UNASSIGNED  # <reserved>..<reserved>
A830..A839  ; DISALLOWED  # NORTH INDIC FRACTION ONE QUARTER..NORTH INDI
A83A..A83F  ; UNASSIGNED  # <reserved>..<reserved>
A840..A873  ; PVALID      # PHAGS-PA LETTER KA..PHAGS-PA LETTER CANDRABI
A874..A877  ; DISALLOWED  # PHAGS-PA SINGLE HEAD MARK..PHAGS-PA MARK DOU
A878..A87F  ; UNASSIGNED  # <reserved>..<reserved>
A880..A8C5  ; PVALID      # SAURASHTRA SIGN ANUSVARA..SAURASHTRA SIGN CA
A8C6..A8CD  ; UNASSIGNED  # <reserved>..<reserved>
A8CE..A8CF  ; DISALLOWED  # SAURASHTRA DANDA..SAURASHTRA DOUBLE DANDA
A8D0..A8D9  ; PVALID      # SAURASHTRA DIGIT ZERO..SAURASHTRA DIGIT NINE
A8DA..A8DF  ; UNASSIGNED  # <reserved>..<reserved>
A8E0..A8F7  ; PVALID      # COMBINING DEVANAGARI DIGIT ZERO..DEVANAGARI
A8F8..A8FA  ; DISALLOWED  # DEVANAGARI SIGN PUSHPIKA..DEVANAGARI CARET
A8FB        ; PVALID      # DEVANAGARI HEADSTROKE
A8FC        ; DISALLOWED  # DEVANAGARI SIGN SIDDHAM
A8FD..A92D  ; PVALID      # DEVANAGARI JAIN OM..KAYAH LI TONE CALYA PLOP
A92E..A92F  ; DISALLOWED  # KAYAH LI SIGN CWI..KAYAH LI SIGN SHYA
A930..A953  ; PVALID      # REJANG LETTER KA..REJANG VIRAMA
A954..A95E  ; UNASSIGNED  # <reserved>..<reserved>
A95F..A97C  ; DISALLOWED  # REJANG SECTION MARK..HANGUL CHOSEONG SSANGYE
A97D..A97F  ; UNASSIGNED  # <reserved>..<reserved>
A980..A9C0  ; PVALID      # JAVANESE SIGN PANYANGGA..JAVANESE PANGKON
A9C1..A9CD  ; DISALLOWED  # JAVANESE LEFT RERENGGAN..JAVANESE TURNED PAD
A9CE        ; UNASSIGNED  # <reserved>
A9CF..A9D9  ; PVALID      # JAVANESE PANGRANGKEP..JAVANESE DIGIT NINE
A9DA..A9DD  ; UNASSIGNED  # <reserved>..<reserved>
A9DE..A9DF  ; DISALLOWED  # JAVANESE PADA TIRTA TUMETES..JAVANESE PADA I
```

```
A9E0..A9FE  ; PVALID      # MYANMAR LETTER SHAN GHA..MYANMAR LETTER TAI
A9FF        ; UNASSIGNED  # <reserved>
AA00..AA36  ; PVALID      # CHAM LETTER A..CHAM CONSONANT SIGN WA
AA37..AA3F  ; UNASSIGNED  # <reserved>..<reserved>
AA40..AA4D  ; PVALID      # CHAM LETTER FINAL K..CHAM CONSONANT SIGN FIN
AA4E..AA4F  ; UNASSIGNED  # <reserved>..<reserved>
AA50..AA59  ; PVALID      # CHAM DIGIT ZERO..CHAM DIGIT NINE
AA5A..AA5B  ; UNASSIGNED  # <reserved>..<reserved>
AA5C..AA5F  ; DISALLOWED  # CHAM PUNCTUATION SPIRAL..CHAM PUNCTUATION TR
AA60..AA76  ; PVALID      # MYANMAR LETTER KHAMTI GA..MYANMAR LOGOGRAM K
AA77..AA79  ; DISALLOWED  # MYANMAR SYMBOL AITON EXCLAMATION..MYANMAR SY
AA7A..AAC2  ; PVALID      # MYANMAR LETTER AITON RA..TAI VIET TONE MAI S
AAC3..AADA  ; UNASSIGNED  # <reserved>..<reserved>
AADB..AADD  ; PVALID      # TAI VIET SYMBOL KON..TAI VIET SYMBOL SAM
AADE..AADF  ; DISALLOWED  # TAI VIET SYMBOL HO HOI..TAI VIET SYMBOL KOI
AAE0..AAEF  ; PVALID      # MEETEI MAYEK LETTER E..MEETEI MAYEK VOWEL SI
AAF0..AAF1  ; DISALLOWED  # MEETEI MAYEK CHEIKHAN..MEETEI MAYEK AHANG KH
AAF2..AAF6  ; PVALID      # MEETEI MAYEK ANJI..MEETEI MAYEK VIRAMA
AAF7..AB00  ; UNASSIGNED  # <reserved>..<reserved>
AB01..AB06  ; PVALID      # ETHIOPIC SYLLABLE TTHU..ETHIOPIC SYLLABLE TT
AB07..AB08  ; UNASSIGNED  # <reserved>..<reserved>
AB09..AB0E  ; PVALID      # ETHIOPIC SYLLABLE DDHU..ETHIOPIC SYLLABLE DD
AB0F..AB10  ; UNASSIGNED  # <reserved>..<reserved>
AB11..AB16  ; PVALID      # ETHIOPIC SYLLABLE DZU..ETHIOPIC SYLLABLE DZO
AB17..AB1F  ; UNASSIGNED  # <reserved>..<reserved>
AB20..AB26  ; PVALID      # ETHIOPIC SYLLABLE CCHHA..ETHIOPIC SYLLABLE C
AB27        ; UNASSIGNED  # <reserved>
AB28..AB2E  ; PVALID      # ETHIOPIC SYLLABLE BBA..ETHIOPIC SYLLABLE BBO
AB2F        ; UNASSIGNED  # <reserved>
AB30..AB5A  ; PVALID      # LATIN SMALL LETTER BARRED ALPHA..LATIN SMALL
AB5B..AB5F  ; DISALLOWED  # MODIFIER BREVE WITH INVERTED BREVE..MODIFIER
AB60..AB65  ; PVALID      # LATIN SMALL LETTER SAKHA YAT..GREEK LETTER S
AB66..AB6F  ; UNASSIGNED  # <reserved>..<reserved>
AB70..ABBF  ; DISALLOWED  # CHEROKEE SMALL LETTER A..CHEROKEE SMALL LETT
ABC0..ABEA  ; PVALID      # MEETEI MAYEK LETTER KOK..MEETEI MAYEK VOWEL
ABEB        ; DISALLOWED  # MEETEI MAYEK CHEIKHEI
ABEC..ABED  ; PVALID      # MEETEI MAYEK LUM IYEK..MEETEI MAYEK APUN IYE
ABEE..ABEF  ; UNASSIGNED  # <reserved>..<reserved>
ABF0..ABF9  ; PVALID      # MEETEI MAYEK DIGIT ZERO..MEETEI MAYEK DIGIT
ABFA..ABFF  ; UNASSIGNED  # <reserved>..<reserved>
AC00..D7A3  ; PVALID      # <Hangul Syllable>..<Hangul Syllable>
D7A4..D7AF  ; UNASSIGNED  # <reserved>..<reserved>
D7B0..D7C6  ; DISALLOWED  # HANGUL JUNGSEONG O-YEO..HANGUL JUNGSEONG ARA
D7C7..D7CA  ; UNASSIGNED  # <reserved>..<reserved>
D7CB..D7FB  ; DISALLOWED  # HANGUL JONGSEONG NIEUN-RIEUL..HANGUL JONGSEO
D7FC..D7FF  ; UNASSIGNED  # <reserved>..<reserved>
D800..FA0D  ; DISALLOWED  # <Non Private Use High Surrogate>..CJK COMPAT
FA0E..FA0F  ; PVALID      # CJK COMPATIBILITY IDEOGRAPH-FA0E..CJK COMPAT
```

```
FA10        ; DISALLOWED  # CJK COMPATIBILITY IDEOGRAPH-FA10
FA11        ; PVALID      # CJK COMPATIBILITY IDEOGRAPH-FA11
FA12        ; DISALLOWED  # CJK COMPATIBILITY IDEOGRAPH-FA12
FA13..FA14  ; PVALID      # CJK COMPATIBILITY IDEOGRAPH-FA13..CJK COMPAT
FA15..FA1E  ; DISALLOWED  # CJK COMPATIBILITY IDEOGRAPH-FA15..CJK COMPAT
FA1F        ; PVALID      # CJK COMPATIBILITY IDEOGRAPH-FA1F
FA20        ; DISALLOWED  # CJK COMPATIBILITY IDEOGRAPH-FA20
FA21        ; PVALID      # CJK COMPATIBILITY IDEOGRAPH-FA21
FA22        ; DISALLOWED  # CJK COMPATIBILITY IDEOGRAPH-FA22
FA23..FA24  ; PVALID      # CJK COMPATIBILITY IDEOGRAPH-FA23..CJK COMPAT
FA25..FA26  ; DISALLOWED  # CJK COMPATIBILITY IDEOGRAPH-FA25..CJK COMPAT
FA27..FA29  ; PVALID      # CJK COMPATIBILITY IDEOGRAPH-FA27..CJK COMPAT
FA2A..FA6D  ; DISALLOWED  # CJK COMPATIBILITY IDEOGRAPH-FA2A..CJK COMPAT
FA6E..FA6F  ; UNASSIGNED  # <reserved>..<reserved>
FA70..FAD9  ; DISALLOWED  # CJK COMPATIBILITY IDEOGRAPH-FA70..CJK COMPAT
FADA..FAFF  ; UNASSIGNED  # <reserved>..<reserved>
FB00..FB06  ; DISALLOWED  # LATIN SMALL LIGATURE FF..LATIN SMALL LIGATUR
FB07..FB12  ; UNASSIGNED  # <reserved>..<reserved>
FB13..FB17  ; DISALLOWED  # ARMENIAN SMALL LIGATURE MEN NOW..ARMENIAN SM
FB18..FB1C  ; UNASSIGNED  # <reserved>..<reserved>
FB1D        ; DISALLOWED  # HEBREW LETTER YOD WITH HIRIQ
FB1E        ; PVALID      # HEBREW POINT JUDEO-SPANISH VARIKA
FB1F..FB36  ; DISALLOWED  # HEBREW LIGATURE YIDDISH YOD YOD PATAH..HEBRE
FB37        ; UNASSIGNED  # <reserved>
FB38..FB3C  ; DISALLOWED  # HEBREW LETTER TET WITH DAGESH..HEBREW LETTER
FB3D        ; UNASSIGNED  # <reserved>
FB3E        ; DISALLOWED  # HEBREW LETTER MEM WITH DAGESH
FB3F        ; UNASSIGNED  # <reserved>
FB40..FB41  ; DISALLOWED  # HEBREW LETTER NUN WITH DAGESH..HEBREW LETTER
FB42        ; UNASSIGNED  # <reserved>
FB43..FB44  ; DISALLOWED  # HEBREW LETTER FINAL PE WITH DAGESH..HEBREW L
FB45        ; UNASSIGNED  # <reserved>
FB46..FBC1  ; DISALLOWED  # HEBREW LETTER TSADI WITH DAGESH..ARABIC SYMB
FBC2..FBD2  ; UNASSIGNED  # <reserved>..<reserved>
FBD3..FD3F  ; DISALLOWED  # ARABIC LETTER NG ISOLATED FORM..ORNATE RIGHT
FD40..FD4F  ; UNASSIGNED  # <reserved>..<reserved>
FD50..FD8F  ; DISALLOWED  # ARABIC LIGATURE TEH WITH JEEM WITH MEEM INIT
FD90..FD91  ; UNASSIGNED  # <reserved>..<reserved>
FD92..FDC7  ; DISALLOWED  # ARABIC LIGATURE MEEM WITH JEEM WITH KHAH INI
FDC8..FDCF  ; UNASSIGNED  # <reserved>..<reserved>
FDD0..FDFD  ; DISALLOWED  # <noncharacter>..ARABIC LIGATURE BISMILLAH AR
FDFE..FDFF  ; UNASSIGNED  # <reserved>..<reserved>
FE00..FE19  ; DISALLOWED  # VARIATION SELECTOR-1..PRESENTATION FORM FOR
FE1A..FE1F  ; UNASSIGNED  # <reserved>..<reserved>
FE20..FE2F  ; PVALID      # COMBINING LIGATURE LEFT HALF..COMBINING CYRI
FE30..FE52  ; DISALLOWED  # PRESENTATION FORM FOR VERTICAL TWO DOT LEADE
FE53        ; UNASSIGNED  # <reserved>
FE54..FE66  ; DISALLOWED  # SMALL SEMICOLON..SMALL EQUALS SIGN
```

```
FE67        ; UNASSIGNED  # <reserved>
FE68..FE6B  ; DISALLOWED  # SMALL REVERSE SOLIDUS..SMALL COMMERCIAL AT
FE6C..FE6F  ; UNASSIGNED  # <reserved>..<reserved>
FE70..FE72  ; DISALLOWED  # ARABIC FATHATAN ISOLATED FORM..ARABIC DAMMAT
FE73        ; PVALID      # ARABIC TAIL FRAGMENT
FE74        ; DISALLOWED  # ARABIC KASRATAN ISOLATED FORM
FE75        ; UNASSIGNED  # <reserved>
FE76..FEFC  ; DISALLOWED  # ARABIC FATHA ISOLATED FORM..ARABIC LIGATURE
FEFD..FEFE  ; UNASSIGNED  # <reserved>..<reserved>
FEFF        ; DISALLOWED  # ZERO WIDTH NO-BREAK SPACE
FF00        ; UNASSIGNED  # <reserved>
FF01..FFBE  ; DISALLOWED  # FULLWIDTH EXCLAMATION MARK..HALFWIDTH HANGUL
FFBF..FFC1  ; UNASSIGNED  # <reserved>..<reserved>
FFC2..FFC7  ; DISALLOWED  # HALFWIDTH HANGUL LETTER A..HALFWIDTH HANGUL
FFC8..FFC9  ; UNASSIGNED  # <reserved>..<reserved>
FFCA..FFCF  ; DISALLOWED  # HALFWIDTH HANGUL LETTER YEO..HALFWIDTH HANGU
FFD0..FFD1  ; UNASSIGNED  # <reserved>..<reserved>
FFD2..FFD7  ; DISALLOWED  # HALFWIDTH HANGUL LETTER YO..HALFWIDTH HANGUL
FFD8..FFD9  ; UNASSIGNED  # <reserved>..<reserved>
FFDA..FFDC  ; DISALLOWED  # HALFWIDTH HANGUL LETTER EU..HALFWIDTH HANGUL
FFDD..FFDF  ; UNASSIGNED  # <reserved>..<reserved>
FFE0..FFE6  ; DISALLOWED  # FULLWIDTH CENT SIGN..FULLWIDTH WON SIGN
FFE7        ; UNASSIGNED  # <reserved>
FFE8..FFEE  ; DISALLOWED  # HALFWIDTH FORMS LIGHT VERTICAL..HALFWIDTH WH
FFEF..FFF8  ; UNASSIGNED  # <reserved>..<reserved>
FFF9..FFFF  ; DISALLOWED  # INTERLINEAR ANNOTATION ANCHOR..<noncharacter
10000..1000B; PVALID      # LINEAR B SYLLABLE B008 A..LINEAR B SYLLABLE
1000C       ; UNASSIGNED  # <reserved>
1000D..10026; PVALID      # LINEAR B SYLLABLE B036 JO..LINEAR B SYLLABLE
10027       ; UNASSIGNED  # <reserved>
10028..1003A; PVALID      # LINEAR B SYLLABLE B060 RA..LINEAR B SYLLABLE
1003B       ; UNASSIGNED  # <reserved>
1003C..1003D; PVALID      # LINEAR B SYLLABLE B017 ZA..LINEAR B SYLLABLE
1003E       ; UNASSIGNED  # <reserved>
1003F..1004D; PVALID      # LINEAR B SYLLABLE B020 ZO..LINEAR B SYLLABLE
1004E..1004F; UNASSIGNED  # <reserved>..<reserved>
10050..1005D; PVALID      # LINEAR B SYMBOL B018..LINEAR B SYMBOL B089
1005E..1007F; UNASSIGNED  # <reserved>..<reserved>
10080..100FA; PVALID      # LINEAR B IDEOGRAM B100 MAN..LINEAR B IDEOGRA
100FB..100FF; UNASSIGNED  # <reserved>..<reserved>
10100..10102; DISALLOWED  # AEGEAN WORD SEPARATOR LINE..AEGEAN CHECK MAR
10103..10106; UNASSIGNED  # <reserved>..<reserved>
10107..10133; DISALLOWED  # AEGEAN NUMBER ONE..AEGEAN NUMBER NINETY THOU
10134..10136; UNASSIGNED  # <reserved>..<reserved>
10137..1018E; DISALLOWED  # AEGEAN WEIGHT BASE UNIT..NOMISMA SIGN
1018F       ; UNASSIGNED  # <reserved>
10190..1019B; DISALLOWED  # ROMAN SEXTANS SIGN..ROMAN CENTURIAL SIGN
1019C..1019F; UNASSIGNED  # <reserved>..<reserved>
```

```
101A0       ; DISALLOWED  # GREEK SYMBOL TAU RHO
101A1..101CF; UNASSIGNED  # <reserved>..<reserved>
101D0..101FC; DISALLOWED  # PHAISTOS DISC SIGN PEDESTRIAN..PHAISTOS DISC
101FD       ; PVALID      # PHAISTOS DISC SIGN COMBINING OBLIQUE STROKE
101FE..1027F; UNASSIGNED  # <reserved>..<reserved>
10280..1029C; PVALID      # LYCIAN LETTER A..LYCIAN LETTER X
1029D..1029F; UNASSIGNED  # <reserved>..<reserved>
102A0..102D0; PVALID      # CARIAN LETTER A..CARIAN LETTER UUU3
102D1..102DF; UNASSIGNED  # <reserved>..<reserved>
102E0       ; PVALID      # COPTIC EPACT THOUSANDS MARK
102E1..102FB; DISALLOWED  # COPTIC EPACT DIGIT ONE..COPTIC EPACT NUMBER
102FC..102FF; UNASSIGNED  # <reserved>..<reserved>
10300..1031F; PVALID      # OLD ITALIC LETTER A..OLD ITALIC LETTER ESS
10320..10323; DISALLOWED  # OLD ITALIC NUMERAL ONE..OLD ITALIC NUMERAL F
10324..1032C; UNASSIGNED  # <reserved>..<reserved>
1032D..10340; PVALID      # OLD ITALIC LETTER YE..GOTHIC LETTER PAIRTHRA
10341       ; DISALLOWED  # GOTHIC LETTER NINETY
10342..10349; PVALID      # GOTHIC LETTER RAIDA..GOTHIC LETTER OTHAL
1034A       ; DISALLOWED  # GOTHIC LETTER NINE HUNDRED
1034B..1034F; UNASSIGNED  # <reserved>..<reserved>
10350..1037A; PVALID      # OLD PERMIC LETTER AN..COMBINING OLD PERMIC L
1037B..1037F; UNASSIGNED  # <reserved>..<reserved>
10380..1039D; PVALID      # UGARITIC LETTER ALPA..UGARITIC LETTER SSU
1039E       ; UNASSIGNED  # <reserved>
1039F       ; DISALLOWED  # UGARITIC WORD DIVIDER
103A0..103C3; PVALID      # OLD PERSIAN SIGN A..OLD PERSIAN SIGN HA
103C4..103C7; UNASSIGNED  # <reserved>..<reserved>
103C8..103CF; PVALID      # OLD PERSIAN SIGN AURAMAZDAA..OLD PERSIAN SIG
103D0..103D5; DISALLOWED  # OLD PERSIAN WORD DIVIDER..OLD PERSIAN NUMBER
103D6..103FF; UNASSIGNED  # <reserved>..<reserved>
10400..10427; DISALLOWED  # DESERET CAPITAL LETTER LONG I..DESERET CAPIT
10428..1049D; PVALID      # DESERET SMALL LETTER LONG I..OSMANYA LETTER
1049E..1049F; UNASSIGNED  # <reserved>..<reserved>
104A0..104A9; PVALID      # OSMANYA DIGIT ZERO..OSMANYA DIGIT NINE
104AA..104AF; UNASSIGNED  # <reserved>..<reserved>
104B0..104D3; DISALLOWED  # OSAGE CAPITAL LETTER A..OSAGE CAPITAL LETTER
104D4..104D7; UNASSIGNED  # <reserved>..<reserved>
104D8..104FB; PVALID      # OSAGE SMALL LETTER A..OSAGE SMALL LETTER ZHA
104FC..104FF; UNASSIGNED  # <reserved>..<reserved>
10500..10527; PVALID      # ELBASAN LETTER A..ELBASAN LETTER KHE
10528..1052F; UNASSIGNED  # <reserved>..<reserved>
10530..10563; PVALID      # CAUCASIAN ALBANIAN LETTER ALT..CAUCASIAN ALB
10564..1056E; UNASSIGNED  # <reserved>..<reserved>
1056F       ; DISALLOWED  # CAUCASIAN ALBANIAN CITATION MARK
10570..105FF; UNASSIGNED  # <reserved>..<reserved>
10600..10736; PVALID      # LINEAR A SIGN AB001..LINEAR A SIGN A664
10737..1073F; UNASSIGNED  # <reserved>..<reserved>
10740..10755; PVALID      # LINEAR A SIGN A701 A..LINEAR A SIGN A732 JE
```

```
10756..1075F; UNASSIGNED  # <reserved>..<reserved>
10760..10767; PVALID      # LINEAR A SIGN A800..LINEAR A SIGN A807
10768..107FF; UNASSIGNED  # <reserved>..<reserved>
10800..10805; PVALID      # CYPRIOT SYLLABLE A..CYPRIOT SYLLABLE JA
10806..10807; UNASSIGNED  # <reserved>..<reserved>
10808       ; PVALID      # CYPRIOT SYLLABLE JO
10809       ; UNASSIGNED  # <reserved>
1080A..10835; PVALID      # CYPRIOT SYLLABLE KA..CYPRIOT SYLLABLE WO
10836       ; UNASSIGNED  # <reserved>
10837..10838; PVALID      # CYPRIOT SYLLABLE XA..CYPRIOT SYLLABLE XE
10839..1083B; UNASSIGNED  # <reserved>..<reserved>
1083C       ; PVALID      # CYPRIOT SYLLABLE ZA
1083D..1083E; UNASSIGNED  # <reserved>..<reserved>
1083F..10855; PVALID      # CYPRIOT SYLLABLE ZO..IMPERIAL ARAMAIC LETTER
10856       ; UNASSIGNED  # <reserved>
10857..1085F; DISALLOWED  # IMPERIAL ARAMAIC SECTION SIGN..IMPERIAL ARAM
10860..10876; PVALID      # PALMYRENE LETTER ALEPH..PALMYRENE LETTER TAW
10877..1087F; DISALLOWED  # PALMYRENE LEFT-POINTING FLEURON..PALMYRENE N
10880..1089E; PVALID      # NABATAEAN LETTER FINAL ALEPH..NABATAEAN LETT
1089F..108A6; UNASSIGNED  # <reserved>..<reserved>
108A7..108AF; DISALLOWED  # NABATAEAN NUMBER ONE..NABATAEAN NUMBER ONE H
108B0..108DF; UNASSIGNED  # <reserved>..<reserved>
108E0..108F2; PVALID      # HATRAN LETTER ALEPH..HATRAN LETTER QOPH
108F3       ; UNASSIGNED  # <reserved>
108F4..108F5; PVALID      # HATRAN LETTER SHIN..HATRAN LETTER TAW
108F6..108FA; UNASSIGNED  # <reserved>..<reserved>
108FB..108FF; DISALLOWED  # HATRAN NUMBER ONE..HATRAN NUMBER ONE HUNDRED
10900..10915; PVALID      # PHOENICIAN LETTER ALF..PHOENICIAN LETTER TAU
10916..1091B; DISALLOWED  # PHOENICIAN NUMBER ONE..PHOENICIAN NUMBER THR
1091C..1091E; UNASSIGNED  # <reserved>..<reserved>
1091F       ; DISALLOWED  # PHOENICIAN WORD SEPARATOR
10920..10939; PVALID      # LYDIAN LETTER A..LYDIAN LETTER C
1093A..1093E; UNASSIGNED  # <reserved>..<reserved>
1093F       ; DISALLOWED  # LYDIAN TRIANGULAR MARK
10940..1097F; UNASSIGNED  # <reserved>..<reserved>
10980..109B7; PVALID      # MEROITIC HIEROGLYPHIC LETTER A..MEROITIC CUR
109B8..109BB; UNASSIGNED  # <reserved>..<reserved>
109BC..109BD; DISALLOWED  # MEROITIC CURSIVE FRACTION ELEVEN TWELFTHS..M
109BE..109BF; PVALID      # MEROITIC CURSIVE LOGOGRAM RMT..MEROITIC CURS
109C0..109CF; DISALLOWED  # MEROITIC CURSIVE NUMBER ONE..MEROITIC CURSIV
109D0..109D1; UNASSIGNED  # <reserved>..<reserved>
109D2..109FF; DISALLOWED  # MEROITIC CURSIVE NUMBER ONE HUNDRED..MEROITI
10A00..10A03; PVALID      # KHAROSHTHI LETTER A..KHAROSHTHI VOWEL SIGN V
10A04       ; UNASSIGNED  # <reserved>
10A05..10A06; PVALID      # KHAROSHTHI VOWEL SIGN E..KHAROSHTHI VOWEL SI
10A07..10A0B; UNASSIGNED  # <reserved>..<reserved>
10A0C..10A13; PVALID      # KHAROSHTHI VOWEL LENGTH MARK..KHAROSHTHI LET
10A14       ; UNASSIGNED  # <reserved>
```

```
10A15..10A17; PVALID      # KHAROSHTHI LETTER CA..KHAROSHTHI LETTER JA
10A18       ; UNASSIGNED  # <reserved>
10A19..10A35; PVALID      # KHAROSHTHI LETTER NYA..KHAROSHTHI LETTER VHA
10A36..10A37; UNASSIGNED  # <reserved>..<reserved>
10A38..10A3A; PVALID      # KHAROSHTHI SIGN BAR ABOVE..KHAROSHTHI SIGN D
10A3B..10A3E; UNASSIGNED  # <reserved>..<reserved>
10A3F       ; PVALID      # KHAROSHTHI VIRAMA
10A40..10A48; DISALLOWED  # KHAROSHTHI DIGIT ONE..KHAROSHTHI FRACTION ON
10A49..10A4F; UNASSIGNED  # <reserved>..<reserved>
10A50..10A58; DISALLOWED  # KHAROSHTHI PUNCTUATION DOT..KHAROSHTHI PUNCT
10A59..10A5F; UNASSIGNED  # <reserved>..<reserved>
10A60..10A7C; PVALID      # OLD SOUTH ARABIAN LETTER HE..OLD SOUTH ARABI
10A7D..10A7F; DISALLOWED  # OLD SOUTH ARABIAN NUMBER ONE..OLD SOUTH ARAB
10A80..10A9C; PVALID      # OLD NORTH ARABIAN LETTER HEH..OLD NORTH ARAB
10A9D..10A9F; DISALLOWED  # OLD NORTH ARABIAN NUMBER ONE..OLD NORTH ARAB
10AA0..10ABF; UNASSIGNED  # <reserved>..<reserved>
10AC0..10AC7; PVALID      # MANICHAEAN LETTER ALEPH..MANICHAEAN LETTER W
10AC8       ; DISALLOWED  # MANICHAEAN SIGN UD
10AC9..10AE6; PVALID      # MANICHAEAN LETTER ZAYIN..MANICHAEAN ABBREVIA
10AE7..10AEA; UNASSIGNED  # <reserved>..<reserved>
10AEB..10AF6; DISALLOWED  # MANICHAEAN NUMBER ONE..MANICHAEAN PUNCTUATIO
10AF7..10AFF; UNASSIGNED  # <reserved>..<reserved>
10B00..10B35; PVALID      # AVESTAN LETTER A..AVESTAN LETTER HE
10B36..10B38; UNASSIGNED  # <reserved>..<reserved>
10B39..10B3F; DISALLOWED  # AVESTAN ABBREVIATION MARK..LARGE ONE RING OV
10B40..10B55; PVALID      # INSCRIPTIONAL PARTHIAN LETTER ALEPH..INSCRIP
10B56..10B57; UNASSIGNED  # <reserved>..<reserved>
10B58..10B5F; DISALLOWED  # INSCRIPTIONAL PARTHIAN NUMBER ONE..INSCRIPTI
10B60..10B72; PVALID      # INSCRIPTIONAL PAHLAVI LETTER ALEPH..INSCRIPT
10B73..10B77; UNASSIGNED  # <reserved>..<reserved>
10B78..10B7F; DISALLOWED  # INSCRIPTIONAL PAHLAVI NUMBER ONE..INSCRIPTIO
10B80..10B91; PVALID      # PSALTER PAHLAVI LETTER ALEPH..PSALTER PAHLAV
10B92..10B98; UNASSIGNED  # <reserved>..<reserved>
10B99..10B9C; DISALLOWED  # PSALTER PAHLAVI SECTION MARK..PSALTER PAHLAV
10B9D..10BA8; UNASSIGNED  # <reserved>..<reserved>
10BA9..10BAF; DISALLOWED  # PSALTER PAHLAVI NUMBER ONE..PSALTER PAHLAVI
10BB0..10BFF; UNASSIGNED  # <reserved>..<reserved>
10C00..10C48; PVALID      # OLD TURKIC LETTER ORKHON A..OLD TURKIC LETTE
10C49..10C7F; UNASSIGNED  # <reserved>..<reserved>
10C80..10CB2; DISALLOWED  # OLD HUNGARIAN CAPITAL LETTER A..OLD HUNGARIA
10CB3..10CBF; UNASSIGNED  # <reserved>..<reserved>
10CC0..10CF2; PVALID      # OLD HUNGARIAN SMALL LETTER A..OLD HUNGARIAN
10CF3..10CF9; UNASSIGNED  # <reserved>..<reserved>
10CFA..10CFF; DISALLOWED  # OLD HUNGARIAN NUMBER ONE..OLD HUNGARIAN NUMB
10D00..10D27; PVALID      # HANIFI ROHINGYA LETTER A..HANIFI ROHINGYA SI
10D28..10D2F; UNASSIGNED  # <reserved>..<reserved>
10D30..10D39; PVALID      # HANIFI ROHINGYA DIGIT ZERO..HANIFI ROHINGYA
10D3A..10E5F; UNASSIGNED  # <reserved>..<reserved>
```

```
10E60..10E7E; DISALLOWED  # RUMI DIGIT ONE..RUMI FRACTION TWO THIRDS
10E7F..10EFF; UNASSIGNED  # <reserved>..<reserved>
10F00..10F1C; PVALID      # OLD SOGDIAN LETTER ALEPH..OLD SOGDIAN LETTER
10F1D..10F26; DISALLOWED  # OLD SOGDIAN NUMBER ONE..OLD SOGDIAN FRACTION
10F27       ; PVALID      # OLD SOGDIAN LIGATURE AYIN-DALETH
10F28..10F2F; UNASSIGNED  # <reserved>..<reserved>
10F30..10F50; PVALID      # SOGDIAN LETTER ALEPH..SOGDIAN COMBINING STRO
10F51..10F59; DISALLOWED  # SOGDIAN NUMBER ONE..SOGDIAN PUNCTUATION HALF
10F5A..10FFF; UNASSIGNED  # <reserved>..<reserved>
11000..11046; PVALID      # BRAHMI SIGN CANDRABINDU..BRAHMI VIRAMA
11047..1104D; DISALLOWED  # BRAHMI DANDA..BRAHMI PUNCTUATION LOTUS
1104E..11051; UNASSIGNED  # <reserved>..<reserved>
11052..11065; DISALLOWED  # BRAHMI NUMBER ONE..BRAHMI NUMBER ONE THOUSAN
11066..1106F; PVALID      # BRAHMI DIGIT ZERO..BRAHMI DIGIT NINE
11070..1107E; UNASSIGNED  # <reserved>..<reserved>
1107F..110BA; PVALID      # BRAHMI NUMBER JOINER..KAITHI SIGN NUKTA
110BB..110C1; DISALLOWED  # KAITHI ABBREVIATION SIGN..KAITHI DOUBLE DAND
110C2..110CC; UNASSIGNED  # <reserved>..<reserved>
110CD       ; DISALLOWED  # KAITHI NUMBER SIGN ABOVE
110CE..110CF; UNASSIGNED  # <reserved>..<reserved>
110D0..110E8; PVALID      # SORA SOMPENG LETTER SAH..SORA SOMPENG LETTER
110E9..110EF; UNASSIGNED  # <reserved>..<reserved>
110F0..110F9; PVALID      # SORA SOMPENG DIGIT ZERO..SORA SOMPENG DIGIT
110FA..110FF; UNASSIGNED  # <reserved>..<reserved>
11100..11134; PVALID      # CHAKMA SIGN CANDRABINDU..CHAKMA MAAYYAA
11135       ; UNASSIGNED  # <reserved>
11136..1113F; PVALID      # CHAKMA DIGIT ZERO..CHAKMA DIGIT NINE
11140..11143; DISALLOWED  # CHAKMA SECTION MARK..CHAKMA QUESTION MARK
11144..11146; PVALID      # CHAKMA LETTER LHAA..CHAKMA VOWEL SIGN EI
11147..1114F; UNASSIGNED  # <reserved>..<reserved>
11150..11173; PVALID      # MAHAJANI LETTER A..MAHAJANI SIGN NUKTA
11174..11175; DISALLOWED  # MAHAJANI ABBREVIATION SIGN..MAHAJANI SECTION
11176       ; PVALID      # MAHAJANI LIGATURE SHRI
11177..1117F; UNASSIGNED  # <reserved>..<reserved>
11180..111C4; PVALID      # SHARADA SIGN CANDRABINDU..SHARADA OM
111C5..111C8; DISALLOWED  # SHARADA DANDA..SHARADA SEPARATOR
111C9..111CC; PVALID      # SHARADA SANDHI MARK..SHARADA EXTRA SHORT VOW
111CD       ; DISALLOWED  # SHARADA SUTRA MARK
111CE..111CF; UNASSIGNED  # <reserved>..<reserved>
111D0..111DA; PVALID      # SHARADA DIGIT ZERO..SHARADA EKAM
111DB       ; DISALLOWED  # SHARADA SIGN SIDDHAM
111DC       ; PVALID      # SHARADA HEADSTROKE
111DD..111DF; DISALLOWED  # SHARADA CONTINUATION SIGN..SHARADA SECTION M
111E0       ; UNASSIGNED  # <reserved>
111E1..111F4; DISALLOWED  # SINHALA ARCHAIC DIGIT ONE..SINHALA ARCHAIC N
111F5..111FF; UNASSIGNED  # <reserved>..<reserved>
11200..11211; PVALID      # KHOJKI LETTER A..KHOJKI LETTER JJA
11212       ; UNASSIGNED  # <reserved>
```

```
11213..11237; PVALID      # KHOJKI LETTER NYA..KHOJKI SIGN SHADDA
11238..1123D; DISALLOWED  # KHOJKI DANDA..KHOJKI ABBREVIATION SIGN
1123E       ; PVALID      # KHOJKI SIGN SUKUN
1123F..1127F; UNASSIGNED  # <reserved>..<reserved>
11280..11286; PVALID      # MULTANI LETTER A..MULTANI LETTER GA
11287       ; UNASSIGNED  # <reserved>
11288       ; PVALID      # MULTANI LETTER GHA
11289       ; UNASSIGNED  # <reserved>
1128A..1128D; PVALID      # MULTANI LETTER CA..MULTANI LETTER JJA
1128E       ; UNASSIGNED  # <reserved>
1128F..1129D; PVALID      # MULTANI LETTER NYA..MULTANI LETTER BA
1129E       ; UNASSIGNED  # <reserved>
1129F..112A8; PVALID      # MULTANI LETTER BHA..MULTANI LETTER RHA
112A9       ; DISALLOWED  # MULTANI SECTION MARK
112AA..112AF; UNASSIGNED  # <reserved>..<reserved>
112B0..112EA; PVALID      # KHUDAWADI LETTER A..KHUDAWADI SIGN VIRAMA
112EB..112EF; UNASSIGNED  # <reserved>..<reserved>
112F0..112F9; PVALID      # KHUDAWADI DIGIT ZERO..KHUDAWADI DIGIT NINE
112FA..112FF; UNASSIGNED  # <reserved>..<reserved>
11300..11303; PVALID      # GRANTHA SIGN COMBINING ANUSVARA ABOVE..GRANT
11304       ; UNASSIGNED  # <reserved>
11305..1130C; PVALID      # GRANTHA LETTER A..GRANTHA LETTER VOCALIC L
1130D..1130E; UNASSIGNED  # <reserved>..<reserved>
1130F..11310; PVALID      # GRANTHA LETTER EE..GRANTHA LETTER AI
11311..11312; UNASSIGNED  # <reserved>..<reserved>
11313..11328; PVALID      # GRANTHA LETTER OO..GRANTHA LETTER NA
11329       ; UNASSIGNED  # <reserved>
1132A..11330; PVALID      # GRANTHA LETTER PA..GRANTHA LETTER RA
11331       ; UNASSIGNED  # <reserved>
11332..11333; PVALID      # GRANTHA LETTER LA..GRANTHA LETTER LLA
11334       ; UNASSIGNED  # <reserved>
11335..11339; PVALID      # GRANTHA LETTER VA..GRANTHA LETTER HA
1133A       ; UNASSIGNED  # <reserved>
1133B..11344; PVALID      # COMBINING BINDU BELOW..GRANTHA VOWEL SIGN VO
11345..11346; UNASSIGNED  # <reserved>..<reserved>
11347..11348; PVALID      # GRANTHA VOWEL SIGN EE..GRANTHA VOWEL SIGN AI
11349..1134A; UNASSIGNED  # <reserved>..<reserved>
1134B..1134D; PVALID      # GRANTHA VOWEL SIGN OO..GRANTHA SIGN VIRAMA
1134E..1134F; UNASSIGNED  # <reserved>..<reserved>
11350       ; PVALID      # GRANTHA OM
11351..11356; UNASSIGNED  # <reserved>..<reserved>
11357       ; PVALID      # GRANTHA AU LENGTH MARK
11358..1135C; UNASSIGNED  # <reserved>..<reserved>
1135D..11363; PVALID      # GRANTHA SIGN PLUTA..GRANTHA VOWEL SIGN VOCAL
11364..11365; UNASSIGNED  # <reserved>..<reserved>
11366..1136C; PVALID      # COMBINING GRANTHA DIGIT ZERO..COMBINING GRAN
1136D..1136F; UNASSIGNED  # <reserved>..<reserved>
11370..11374; PVALID      # COMBINING GRANTHA LETTER A..COMBINING GRANTH
```

```
11375..113FF; UNASSIGNED  # <reserved>..<reserved>
11400..1144A; PVALID      # NEWA LETTER A..NEWA SIDDHI
1144B..1144F; DISALLOWED  # NEWA DANDA..NEWA ABBREVIATION SIGN
11450..11459; PVALID      # NEWA DIGIT ZERO..NEWA DIGIT NINE
1145A       ; UNASSIGNED  # <reserved>
1145B       ; DISALLOWED  # NEWA PLACEHOLDER MARK
1145C       ; UNASSIGNED  # <reserved>
1145D       ; DISALLOWED  # NEWA INSERTION SIGN
1145E       ; PVALID      # NEWA SANDHI MARK
1145F..1147F; UNASSIGNED  # <reserved>..<reserved>
11480..114C5; PVALID      # TIRHUTA ANJI..TIRHUTA GVANG
114C6       ; DISALLOWED  # TIRHUTA ABBREVIATION SIGN
114C7       ; PVALID      # TIRHUTA OM
114C8..114CF; UNASSIGNED  # <reserved>..<reserved>
114D0..114D9; PVALID      # TIRHUTA DIGIT ZERO..TIRHUTA DIGIT NINE
114DA..1157F; UNASSIGNED  # <reserved>..<reserved>
11580..115B5; PVALID      # SIDDHAM LETTER A..SIDDHAM VOWEL SIGN VOCALIC
115B6..115B7; UNASSIGNED  # <reserved>..<reserved>
115B8..115C0; PVALID      # SIDDHAM VOWEL SIGN E..SIDDHAM SIGN NUKTA
115C1..115D7; DISALLOWED  # SIDDHAM SIGN SIDDHAM..SIDDHAM SECTION MARK W
115D8..115DD; PVALID      # SIDDHAM LETTER THREE-CIRCLE ALTERNATE I..SID
115DE..115FF; UNASSIGNED  # <reserved>..<reserved>
11600..11640; PVALID      # MODI LETTER A..MODI SIGN ARDHACANDRA
11641..11643; DISALLOWED  # MODI DANDA..MODI ABBREVIATION SIGN
11644       ; PVALID      # MODI SIGN HUVA
11645..1164F; UNASSIGNED  # <reserved>..<reserved>
11650..11659; PVALID      # MODI DIGIT ZERO..MODI DIGIT NINE
1165A..1165F; UNASSIGNED  # <reserved>..<reserved>
11660..1166C; DISALLOWED  # MONGOLIAN BIRGA WITH ORNAMENT..MONGOLIAN TUR
1166D..1167F; UNASSIGNED  # <reserved>..<reserved>
11680..116B7; PVALID      # TAKRI LETTER A..TAKRI SIGN NUKTA
116B8..116BF; UNASSIGNED  # <reserved>..<reserved>
116C0..116C9; PVALID      # TAKRI DIGIT ZERO..TAKRI DIGIT NINE
116CA..116FF; UNASSIGNED  # <reserved>..<reserved>
11700..1171A; PVALID      # AHOM LETTER KA..AHOM LETTER ALTERNATE BA
1171B..1171C; UNASSIGNED  # <reserved>..<reserved>
1171D..1172B; PVALID      # AHOM CONSONANT SIGN MEDIAL LA..AHOM SIGN KIL
1172C..1172F; UNASSIGNED  # <reserved>..<reserved>
11730..11739; PVALID      # AHOM DIGIT ZERO..AHOM DIGIT NINE
1173A..1173F; DISALLOWED  # AHOM NUMBER TEN..AHOM SYMBOL VI
11740..117FF; UNASSIGNED  # <reserved>..<reserved>
11800..1183A; PVALID      # DOGRA LETTER A..DOGRA SIGN NUKTA
1183B       ; DISALLOWED  # DOGRA ABBREVIATION SIGN
1183C..1189F; UNASSIGNED  # <reserved>..<reserved>
118A0..118BF; DISALLOWED  # WARANG CITI CAPITAL LETTER NGAA..WARANG CITI
118C0..118E9; PVALID      # WARANG CITI SMALL LETTER NGAA..WARANG CITI D
118EA..118F2; DISALLOWED  # WARANG CITI NUMBER TEN..WARANG CITI NUMBER N
118F3..118FE; UNASSIGNED  # <reserved>..<reserved>
```

```
118FF       ; PVALID      # WARANG CITI OM
11900..119FF; UNASSIGNED  # <reserved>..<reserved>
11A00..11A3E; PVALID      # ZANABAZAR SQUARE LETTER A..ZANABAZAR SQUARE
11A3F..11A46; DISALLOWED  # ZANABAZAR SQUARE INITIAL HEAD MARK..ZANABAZA
11A47       ; PVALID      # ZANABAZAR SQUARE SUBJOINER
11A48..11A4F; UNASSIGNED  # <reserved>..<reserved>
11A50..11A83; PVALID      # SOYOMBO LETTER A..SOYOMBO LETTER KSSA
11A84..11A85; UNASSIGNED  # <reserved>..<reserved>
11A86..11A99; PVALID      # SOYOMBO CLUSTER-INITIAL LETTER RA..SOYOMBO S
11A9A..11A9C; DISALLOWED  # SOYOMBO MARK TSHEG..SOYOMBO MARK DOUBLE SHAD
11A9D       ; PVALID      # SOYOMBO MARK PLUTA
11A9E..11AA2; DISALLOWED  # SOYOMBO HEAD MARK WITH MOON AND SUN AND TRIP
11AA3..11ABF; UNASSIGNED  # <reserved>..<reserved>
11AC0..11AF8; PVALID      # PAU CIN HAU LETTER PA..PAU CIN HAU GLOTTAL S
11AF9..11BFF; UNASSIGNED  # <reserved>..<reserved>
11C00..11C08; PVALID      # BHAIKSUKI LETTER A..BHAIKSUKI LETTER VOCALIC
11C09       ; UNASSIGNED  # <reserved>
11C0A..11C36; PVALID      # BHAIKSUKI LETTER E..BHAIKSUKI VOWEL SIGN VOC
11C37       ; UNASSIGNED  # <reserved>
11C38..11C40; PVALID      # BHAIKSUKI VOWEL SIGN E..BHAIKSUKI SIGN AVAGR
11C41..11C45; DISALLOWED  # BHAIKSUKI DANDA..BHAIKSUKI GAP FILLER-2
11C46..11C4F; UNASSIGNED  # <reserved>..<reserved>
11C50..11C59; PVALID      # BHAIKSUKI DIGIT ZERO..BHAIKSUKI DIGIT NINE
11C5A..11C6C; DISALLOWED  # BHAIKSUKI NUMBER ONE..BHAIKSUKI HUNDREDS UNI
11C6D..11C6F; UNASSIGNED  # <reserved>..<reserved>
11C70..11C71; DISALLOWED  # MARCHEN HEAD MARK..MARCHEN MARK SHAD
11C72..11C8F; PVALID      # MARCHEN LETTER KA..MARCHEN LETTER A
11C90..11C91; UNASSIGNED  # <reserved>..<reserved>
11C92..11CA7; PVALID      # MARCHEN SUBJOINED LETTER KA..MARCHEN SUBJOIN
11CA8       ; UNASSIGNED  # <reserved>
11CA9..11CB6; PVALID      # MARCHEN SUBJOINED LETTER YA..MARCHEN SIGN CA
11CB7..11CFF; UNASSIGNED  # <reserved>..<reserved>
11D00..11D06; PVALID      # MASARAM GONDI LETTER A..MASARAM GONDI LETTER
11D07       ; UNASSIGNED  # <reserved>
11D08..11D09; PVALID      # MASARAM GONDI LETTER AI..MASARAM GONDI LETTE
11D0A       ; UNASSIGNED  # <reserved>
11D0B..11D36; PVALID      # MASARAM GONDI LETTER AU..MASARAM GONDI VOWEL
11D37..11D39; UNASSIGNED  # <reserved>..<reserved>
11D3A       ; PVALID      # MASARAM GONDI VOWEL SIGN E
11D3B       ; UNASSIGNED  # <reserved>
11D3C..11D3D; PVALID      # MASARAM GONDI VOWEL SIGN AI..MASARAM GONDI V
11D3E       ; UNASSIGNED  # <reserved>
11D3F..11D47; PVALID      # MASARAM GONDI VOWEL SIGN AU..MASARAM GONDI R
11D48..11D4F; UNASSIGNED  # <reserved>..<reserved>
11D50..11D59; PVALID      # MASARAM GONDI DIGIT ZERO..MASARAM GONDI DIGI
11D5A..11D5F; UNASSIGNED  # <reserved>..<reserved>
11D60..11D65; PVALID      # GUNJALA GONDI LETTER A..GUNJALA GONDI LETTER
11D66       ; UNASSIGNED  # <reserved>
```

```
11D67..11D68; PVALID      # GUNJALA GONDI LETTER EE..GUNJALA GONDI LETTE
11D69       ; UNASSIGNED  # <reserved>
11D6A..11D8E; PVALID      # GUNJALA GONDI LETTER OO..GUNJALA GONDI VOWEL
11D8F       ; UNASSIGNED  # <reserved>
11D90..11D91; PVALID      # GUNJALA GONDI VOWEL SIGN EE..GUNJALA GONDI V
11D92       ; UNASSIGNED  # <reserved>
11D93..11D98; PVALID      # GUNJALA GONDI VOWEL SIGN OO..GUNJALA GONDI O
11D99..11D9F; UNASSIGNED  # <reserved>..<reserved>
11DA0..11DA9; PVALID      # GUNJALA GONDI DIGIT ZERO..GUNJALA GONDI DIGI
11DAA..11EDF; UNASSIGNED  # <reserved>..<reserved>
11EE0..11EF6; PVALID      # MAKASAR LETTER KA..MAKASAR VOWEL SIGN O
11EF7..11EF8; DISALLOWED  # MAKASAR PASSIMBANG..MAKASAR END OF SECTION
11EF9..11FFF; UNASSIGNED  # <reserved>..<reserved>
12000..12399; PVALID      # CUNEIFORM SIGN A..CUNEIFORM SIGN U U
1239A..123FF; UNASSIGNED  # <reserved>..<reserved>
12400..1246E; DISALLOWED  # CUNEIFORM NUMERIC SIGN TWO ASH..CUNEIFORM NU
1246F       ; UNASSIGNED  # <reserved>
12470..12474; DISALLOWED  # CUNEIFORM PUNCTUATION SIGN OLD ASSYRIAN WORD
12475..1247F; UNASSIGNED  # <reserved>..<reserved>
12480..12543; PVALID      # CUNEIFORM SIGN AB TIMES NUN TENU..CUNEIFORM
12544..12FFF; UNASSIGNED  # <reserved>..<reserved>
13000..1342E; PVALID      # EGYPTIAN HIEROGLYPH A001..EGYPTIAN HIEROGLYP
1342F..143FF; UNASSIGNED  # <reserved>..<reserved>
14400..14646; PVALID      # ANATOLIAN HIEROGLYPH A001..ANATOLIAN HIEROGL
14647..167FF; UNASSIGNED  # <reserved>..<reserved>
16800..16A38; PVALID      # BAMUM LETTER PHASE-A NGKUE MFON..BAMUM LETTE
16A39..16A3F; UNASSIGNED  # <reserved>..<reserved>
16A40..16A5E; PVALID      # MRO LETTER TA..MRO LETTER TEK
16A5F       ; UNASSIGNED  # <reserved>
16A60..16A69; PVALID      # MRO DIGIT ZERO..MRO DIGIT NINE
16A6A..16A6D; UNASSIGNED  # <reserved>..<reserved>
16A6E..16A6F; DISALLOWED  # MRO DANDA..MRO DOUBLE DANDA
16A70..16ACF; UNASSIGNED  # <reserved>..<reserved>
16AD0..16AED; PVALID      # BASSA VAH LETTER ENNI..BASSA VAH LETTER I
16AEE..16AEF; UNASSIGNED  # <reserved>..<reserved>
16AF0..16AF4; PVALID      # BASSA VAH COMBINING HIGH TONE..BASSA VAH COM
16AF5       ; DISALLOWED  # BASSA VAH FULL STOP
16AF6..16AFF; UNASSIGNED  # <reserved>..<reserved>
16B00..16B36; PVALID      # PAHAWH HMONG VOWEL KEEB..PAHAWH HMONG MARK C
16B37..16B3F; DISALLOWED  # PAHAWH HMONG SIGN VOS THOM..PAHAWH HMONG SIG
16B40..16B43; PVALID      # PAHAWH HMONG SIGN VOS SEEV..PAHAWH HMONG SIG
16B44..16B45; DISALLOWED  # PAHAWH HMONG SIGN XAUS..PAHAWH HMONG SIGN CI
16B46..16B4F; UNASSIGNED  # <reserved>..<reserved>
16B50..16B59; PVALID      # PAHAWH HMONG DIGIT ZERO..PAHAWH HMONG DIGIT
16B5A       ; UNASSIGNED  # <reserved>
16B5B..16B61; DISALLOWED  # PAHAWH HMONG NUMBER TENS..PAHAWH HMONG NUMBE
16B62       ; UNASSIGNED  # <reserved>
16B63..16B77; PVALID      # PAHAWH HMONG SIGN VOS LUB..PAHAWH HMONG SIGN
```

```
16B78..16B7C; UNASSIGNED  # <reserved>..<reserved>
16B7D..16B8F; PVALID      # PAHAWH HMONG CLAN SIGN TSHEEJ..PAHAWH HMONG
16B90..16E3F; UNASSIGNED  # <reserved>..<reserved>
16E40..16E5F; DISALLOWED  # MEDEFAIDRIN CAPITAL LETTER M..MEDEFAIDRIN CA
16E60..16E7F; PVALID      # MEDEFAIDRIN SMALL LETTER M..MEDEFAIDRIN SMAL
16E80..16E9A; DISALLOWED  # MEDEFAIDRIN DIGIT ZERO..MEDEFAIDRIN EXCLAMAT
16E9B..16EFF; UNASSIGNED  # <reserved>..<reserved>
16F00..16F44; PVALID      # MIAO LETTER PA..MIAO LETTER HHA
16F45..16F4F; UNASSIGNED  # <reserved>..<reserved>
16F50..16F7E; PVALID      # MIAO LETTER NASALIZATION..MIAO VOWEL SIGN NG
16F7F..16F8E; UNASSIGNED  # <reserved>..<reserved>
16F8F..16F9F; PVALID      # MIAO TONE RIGHT..MIAO LETTER REFORMED TONE-8
16FA0..16FDF; UNASSIGNED  # <reserved>..<reserved>
16FE0..16FE1; PVALID      # TANGUT ITERATION MARK..NUSHU ITERATION MARK
16FE2..16FFF; UNASSIGNED  # <reserved>..<reserved>
17000..187F1; PVALID      # <Tangut Ideograph>..<Tangut Ideograph>
187F2..187FF; UNASSIGNED  # <reserved>..<reserved>
18800..18AF2; PVALID      # TANGUT COMPONENT-001..TANGUT COMPONENT-755
18AF3..1AFFF; UNASSIGNED  # <reserved>..<reserved>
1B000..1B11E; PVALID      # KATAKANA LETTER ARCHAIC E..HENTAIGANA LETTER
1B11F..1B16F; UNASSIGNED  # <reserved>..<reserved>
1B170..1B2FB; PVALID      # NUSHU CHARACTER-1B170..NUSHU CHARACTER-1B2FB
1B2FC..1BBFF; UNASSIGNED  # <reserved>..<reserved>
1BC00..1BC6A; PVALID      # DUPLOYAN LETTER H..DUPLOYAN LETTER VOCALIC M
1BC6B..1BC6F; UNASSIGNED  # <reserved>..<reserved>
1BC70..1BC7C; PVALID      # DUPLOYAN AFFIX LEFT HORIZONTAL SECANT..DUPLO
1BC7D..1BC7F; UNASSIGNED  # <reserved>..<reserved>
1BC80..1BC88; PVALID      # DUPLOYAN AFFIX HIGH ACUTE..DUPLOYAN AFFIX HI
1BC89..1BC8F; UNASSIGNED  # <reserved>..<reserved>
1BC90..1BC99; PVALID      # DUPLOYAN AFFIX LOW ACUTE..DUPLOYAN AFFIX LOW
1BC9A..1BC9B; UNASSIGNED  # <reserved>..<reserved>
1BC9C       ; DISALLOWED  # DUPLOYAN SIGN O WITH CROSS
1BC9D..1BC9E; PVALID      # DUPLOYAN THICK LETTER SELECTOR..DUPLOYAN DOU
1BC9F..1BCA3; DISALLOWED  # DUPLOYAN PUNCTUATION CHINOOK FULL STOP..SHOR
1BCA4..1CFFF; UNASSIGNED  # <reserved>..<reserved>
1D000..1D0F5; DISALLOWED  # BYZANTINE MUSICAL SYMBOL PSILI..BYZANTINE MU
1D0F6..1D0FF; UNASSIGNED  # <reserved>..<reserved>
1D100..1D126; DISALLOWED  # MUSICAL SYMBOL SINGLE BARLINE..MUSICAL SYMBO
1D127..1D128; UNASSIGNED  # <reserved>..<reserved>
1D129..1D1E8; DISALLOWED  # MUSICAL SYMBOL MULTIPLE MEASURE REST..MUSICA
1D1E9..1D1FF; UNASSIGNED  # <reserved>..<reserved>
1D200..1D245; DISALLOWED  # GREEK VOCAL NOTATION SYMBOL-1..GREEK MUSICAL
1D246..1D2DF; UNASSIGNED  # <reserved>..<reserved>
1D2E0..1D2F3; DISALLOWED  # MAYAN NUMERAL ZERO..MAYAN NUMERAL NINETEEN
1D2F4..1D2FF; UNASSIGNED  # <reserved>..<reserved>
1D300..1D356; DISALLOWED  # MONOGRAM FOR EARTH..TETRAGRAM FOR FOSTERING
1D357..1D35F; UNASSIGNED  # <reserved>..<reserved>
1D360..1D378; DISALLOWED  # COUNTING ROD UNIT DIGIT ONE..TALLY MARK FIVE
```

```
1D379..1D3FF; UNASSIGNED  # <reserved>..<reserved>
1D400..1D454; DISALLOWED  # MATHEMATICAL BOLD CAPITAL A..MATHEMATICAL IT
1D455       ; UNASSIGNED  # <reserved>
1D456..1D49C; DISALLOWED  # MATHEMATICAL ITALIC SMALL I..MATHEMATICAL SC
1D49D       ; UNASSIGNED  # <reserved>
1D49E..1D49F; DISALLOWED  # MATHEMATICAL SCRIPT CAPITAL C..MATHEMATICAL
1D4A0..1D4A1; UNASSIGNED  # <reserved>..<reserved>
1D4A2       ; DISALLOWED  # MATHEMATICAL SCRIPT CAPITAL G
1D4A3..1D4A4; UNASSIGNED  # <reserved>..<reserved>
1D4A5..1D4A6; DISALLOWED  # MATHEMATICAL SCRIPT CAPITAL J..MATHEMATICAL
1D4A7..1D4A8; UNASSIGNED  # <reserved>..<reserved>
1D4A9..1D4AC; DISALLOWED  # MATHEMATICAL SCRIPT CAPITAL N..MATHEMATICAL
1D4AD       ; UNASSIGNED  # <reserved>
1D4AE..1D4B9; DISALLOWED  # MATHEMATICAL SCRIPT CAPITAL S..MATHEMATICAL
1D4BA       ; UNASSIGNED  # <reserved>
1D4BB       ; DISALLOWED  # MATHEMATICAL SCRIPT SMALL F
1D4BC       ; UNASSIGNED  # <reserved>
1D4BD..1D4C3; DISALLOWED  # MATHEMATICAL SCRIPT SMALL H..MATHEMATICAL SC
1D4C4       ; UNASSIGNED  # <reserved>
1D4C5..1D505; DISALLOWED  # MATHEMATICAL SCRIPT SMALL P..MATHEMATICAL FR
1D506       ; UNASSIGNED  # <reserved>
1D507..1D50A; DISALLOWED  # MATHEMATICAL FRAKTUR CAPITAL D..MATHEMATICAL
1D50B..1D50C; UNASSIGNED  # <reserved>..<reserved>
1D50D..1D514; DISALLOWED  # MATHEMATICAL FRAKTUR CAPITAL J..MATHEMATICAL
1D515       ; UNASSIGNED  # <reserved>
1D516..1D51C; DISALLOWED  # MATHEMATICAL FRAKTUR CAPITAL S..MATHEMATICAL
1D51D       ; UNASSIGNED  # <reserved>
1D51E..1D539; DISALLOWED  # MATHEMATICAL FRAKTUR SMALL A..MATHEMATICAL D
1D53A       ; UNASSIGNED  # <reserved>
1D53B..1D53E; DISALLOWED  # MATHEMATICAL DOUBLE-STRUCK CAPITAL D..MATHEM
1D53F       ; UNASSIGNED  # <reserved>
1D540..1D544; DISALLOWED  # MATHEMATICAL DOUBLE-STRUCK CAPITAL I..MATHEM
1D545       ; UNASSIGNED  # <reserved>
1D546       ; DISALLOWED  # MATHEMATICAL DOUBLE-STRUCK CAPITAL O
1D547..1D549; UNASSIGNED  # <reserved>..<reserved>
1D54A..1D550; DISALLOWED  # MATHEMATICAL DOUBLE-STRUCK CAPITAL S..MATHEM
1D551       ; UNASSIGNED  # <reserved>
1D552..1D6A5; DISALLOWED  # MATHEMATICAL DOUBLE-STRUCK SMALL A..MATHEMAT
1D6A6..1D6A7; UNASSIGNED  # <reserved>..<reserved>
1D6A8..1D7CB; DISALLOWED  # MATHEMATICAL BOLD CAPITAL ALPHA..MATHEMATICA
1D7CC..1D7CD; UNASSIGNED  # <reserved>..<reserved>
1D7CE..1D9FF; DISALLOWED  # MATHEMATICAL BOLD DIGIT ZERO..SIGNWRITING HE
1DA00..1DA36; PVALID      # SIGNWRITING HEAD RIM..SIGNWRITING AIR SUCKIN
1DA37..1DA3A; DISALLOWED  # SIGNWRITING AIR BLOW SMALL ROTATIONS..SIGNWR
1DA3B..1DA6C; PVALID      # SIGNWRITING MOUTH CLOSED NEUTRAL..SIGNWRITIN
1DA6D..1DA74; DISALLOWED  # SIGNWRITING SHOULDER HIP SPINE..SIGNWRITING
1DA75       ; PVALID      # SIGNWRITING UPPER BODY TILTING FROM HIP JOIN
1DA76..1DA83; DISALLOWED  # SIGNWRITING LIMB COMBINATION..SIGNWRITING LO
```

```
1DA84       ; PVALID      # SIGNWRITING LOCATION HEAD NECK
1DA85..1DA8B; DISALLOWED  # SIGNWRITING LOCATION TORSO..SIGNWRITING PARE
1DA8C..1DA9A; UNASSIGNED  # <reserved>..<reserved>
1DA9B..1DA9F; PVALID      # SIGNWRITING FILL MODIFIER-2..SIGNWRITING FIL
1DAA0       ; UNASSIGNED  # <reserved>
1DAA1..1DAAF; PVALID      # SIGNWRITING ROTATION MODIFIER-2..SIGNWRITING
1DAB0..1DFFF; UNASSIGNED  # <reserved>..<reserved>
1E000..1E006; PVALID      # COMBINING GLAGOLITIC LETTER AZU..COMBINING G
1E007       ; UNASSIGNED  # <reserved>
1E008..1E018; PVALID      # COMBINING GLAGOLITIC LETTER ZEMLJA..COMBININ
1E019..1E01A; UNASSIGNED  # <reserved>..<reserved>
1E01B..1E021; PVALID      # COMBINING GLAGOLITIC LETTER SHTA..COMBINING
1E022       ; UNASSIGNED  # <reserved>
1E023..1E024; PVALID      # COMBINING GLAGOLITIC LETTER YU..COMBINING GL
1E025       ; UNASSIGNED  # <reserved>
1E026..1E02A; PVALID      # COMBINING GLAGOLITIC LETTER YO..COMBINING GL
1E02B..1E7FF; UNASSIGNED  # <reserved>..<reserved>
1E800..1E8C4; PVALID      # MENDE KIKAKUI SYLLABLE M001 KI..MENDE KIKAKU
1E8C5..1E8C6; UNASSIGNED  # <reserved>..<reserved>
1E8C7..1E8CF; DISALLOWED  # MENDE KIKAKUI DIGIT ONE..MENDE KIKAKUI DIGIT
1E8D0..1E8D6; PVALID      # MENDE KIKAKUI COMBINING NUMBER TEENS..MENDE
1E8D7..1E8FF; UNASSIGNED  # <reserved>..<reserved>
1E900..1E921; DISALLOWED  # ADLAM CAPITAL LETTER ALIF..ADLAM CAPITAL LET
1E922..1E94A; PVALID      # ADLAM SMALL LETTER ALIF..ADLAM NUKTA
1E94B..1E94F; UNASSIGNED  # <reserved>..<reserved>
1E950..1E959; PVALID      # ADLAM DIGIT ZERO..ADLAM DIGIT NINE
1E95A..1E95D; UNASSIGNED  # <reserved>..<reserved>
1E95E..1E95F; DISALLOWED  # ADLAM INITIAL EXCLAMATION MARK..ADLAM INITIA
1E960..1EC70; UNASSIGNED  # <reserved>..<reserved>
1EC71..1ECB4; DISALLOWED  # INDIC SIYAQ NUMBER ONE..INDIC SIYAQ ALTERNAT
1ECB5..1EDFF; UNASSIGNED  # <reserved>..<reserved>
1EE00..1EE03; DISALLOWED  # ARABIC MATHEMATICAL ALEF..ARABIC MATHEMATICA
1EE04       ; UNASSIGNED  # <reserved>
1EE05..1EE1F; DISALLOWED  # ARABIC MATHEMATICAL WAW..ARABIC MATHEMATICAL
1EE20       ; UNASSIGNED  # <reserved>
1EE21..1EE22; DISALLOWED  # ARABIC MATHEMATICAL INITIAL BEH..ARABIC MATH
1EE23       ; UNASSIGNED  # <reserved>
1EE24       ; DISALLOWED  # ARABIC MATHEMATICAL INITIAL HEH
1EE25..1EE26; UNASSIGNED  # <reserved>..<reserved>
1EE27       ; DISALLOWED  # ARABIC MATHEMATICAL INITIAL HAH
1EE28       ; UNASSIGNED  # <reserved>
1EE29..1EE32; DISALLOWED  # ARABIC MATHEMATICAL INITIAL YEH..ARABIC MATH
1EE33       ; UNASSIGNED  # <reserved>
1EE34..1EE37; DISALLOWED  # ARABIC MATHEMATICAL INITIAL SHEEN..ARABIC MA
1EE38       ; UNASSIGNED  # <reserved>
1EE39       ; DISALLOWED  # ARABIC MATHEMATICAL INITIAL DAD
1EE3A       ; UNASSIGNED  # <reserved>
1EE3B       ; DISALLOWED  # ARABIC MATHEMATICAL INITIAL GHAIN
```

```
1EE3C..1EE41; UNASSIGNED  # <reserved>..<reserved>
1EE42       ; DISALLOWED  # ARABIC MATHEMATICAL TAILED JEEM
1EE43..1EE46; UNASSIGNED  # <reserved>..<reserved>
1EE47       ; DISALLOWED  # ARABIC MATHEMATICAL TAILED HAH
1EE48       ; UNASSIGNED  # <reserved>
1EE49       ; DISALLOWED  # ARABIC MATHEMATICAL TAILED YEH
1EE4A       ; UNASSIGNED  # <reserved>
1EE4B       ; DISALLOWED  # ARABIC MATHEMATICAL TAILED LAM
1EE4C       ; UNASSIGNED  # <reserved>
1EE4D..1EE4F; DISALLOWED  # ARABIC MATHEMATICAL TAILED NOON..ARABIC MATH
1EE50       ; UNASSIGNED  # <reserved>
1EE51..1EE52; DISALLOWED  # ARABIC MATHEMATICAL TAILED SAD..ARABIC MATHE
1EE53       ; UNASSIGNED  # <reserved>
1EE54       ; DISALLOWED  # ARABIC MATHEMATICAL TAILED SHEEN
1EE55..1EE56; UNASSIGNED  # <reserved>..<reserved>
1EE57       ; DISALLOWED  # ARABIC MATHEMATICAL TAILED KHAH
1EE58       ; UNASSIGNED  # <reserved>
1EE59       ; DISALLOWED  # ARABIC MATHEMATICAL TAILED DAD
1EE5A       ; UNASSIGNED  # <reserved>
1EE5B       ; DISALLOWED  # ARABIC MATHEMATICAL TAILED GHAIN
1EE5C       ; UNASSIGNED  # <reserved>
1EE5D       ; DISALLOWED  # ARABIC MATHEMATICAL TAILED DOTLESS NOON
1EE5E       ; UNASSIGNED  # <reserved>
1EE5F       ; DISALLOWED  # ARABIC MATHEMATICAL TAILED DOTLESS QAF
1EE60       ; UNASSIGNED  # <reserved>
1EE61..1EE62; DISALLOWED  # ARABIC MATHEMATICAL STRETCHED BEH..ARABIC MA
1EE63       ; UNASSIGNED  # <reserved>
1EE64       ; DISALLOWED  # ARABIC MATHEMATICAL STRETCHED HEH
1EE65..1EE66; UNASSIGNED  # <reserved>..<reserved>
1EE67..1EE6A; DISALLOWED  # ARABIC MATHEMATICAL STRETCHED HAH..ARABIC MA
1EE6B       ; UNASSIGNED  # <reserved>
1EE6C..1EE72; DISALLOWED  # ARABIC MATHEMATICAL STRETCHED MEEM..ARABIC M
1EE73       ; UNASSIGNED  # <reserved>
1EE74..1EE77; DISALLOWED  # ARABIC MATHEMATICAL STRETCHED SHEEN..ARABIC
1EE78       ; UNASSIGNED  # <reserved>
1EE79..1EE7C; DISALLOWED  # ARABIC MATHEMATICAL STRETCHED DAD..ARABIC MA
1EE7D       ; UNASSIGNED  # <reserved>
1EE7E       ; DISALLOWED  # ARABIC MATHEMATICAL STRETCHED DOTLESS FEH
1EE7F       ; UNASSIGNED  # <reserved>
1EE80..1EE89; DISALLOWED  # ARABIC MATHEMATICAL LOOPED ALEF..ARABIC MATH
1EE8A       ; UNASSIGNED  # <reserved>
1EE8B..1EE9B; DISALLOWED  # ARABIC MATHEMATICAL LOOPED LAM..ARABIC MATHE
1EE9C..1EEA0; UNASSIGNED  # <reserved>..<reserved>
1EEA1..1EEA3; DISALLOWED  # ARABIC MATHEMATICAL DOUBLE-STRUCK BEH..ARABI
1EEA4       ; UNASSIGNED  # <reserved>
1EEA5..1EEA9; DISALLOWED  # ARABIC MATHEMATICAL DOUBLE-STRUCK WAW..ARABI
1EEAA       ; UNASSIGNED  # <reserved>
1EEAB..1EEBB; DISALLOWED  # ARABIC MATHEMATICAL DOUBLE-STRUCK LAM..ARABI
```

```
1EEBC..1EEEF; UNASSIGNED  # <reserved>..<reserved>
1EEF0..1EEF1; DISALLOWED  # ARABIC MATHEMATICAL OPERATOR MEEM WITH HAH W
1EEF2..1EFFF; UNASSIGNED  # <reserved>..<reserved>
1F000..1F02B; DISALLOWED  # MAHJONG TILE EAST WIND..MAHJONG TILE BACK
1F02C..1F02F; UNASSIGNED  # <reserved>..<reserved>
1F030..1F093; DISALLOWED  # DOMINO TILE HORIZONTAL BACK..DOMINO TILE VER
1F094..1F09F; UNASSIGNED  # <reserved>..<reserved>
1F0A0..1F0AE; DISALLOWED  # PLAYING CARD BACK..PLAYING CARD KING OF SPAD
1F0AF..1F0B0; UNASSIGNED  # <reserved>..<reserved>
1F0B1..1F0BF; DISALLOWED  # PLAYING CARD ACE OF HEARTS..PLAYING CARD RED
1F0C0       ; UNASSIGNED  # <reserved>
1F0C1..1F0CF; DISALLOWED  # PLAYING CARD ACE OF DIAMONDS..PLAYING CARD B
1F0D0       ; UNASSIGNED  # <reserved>
1F0D1..1F0F5; DISALLOWED  # PLAYING CARD ACE OF CLUBS..PLAYING CARD TRUM
1F0F6..1F0FF; UNASSIGNED  # <reserved>..<reserved>
1F100..1F10C; DISALLOWED  # DIGIT ZERO FULL STOP..DINGBAT NEGATIVE CIRCL
1F10D..1F10F; UNASSIGNED  # <reserved>..<reserved>
1F110..1F16B; DISALLOWED  # PARENTHESIZED LATIN CAPITAL LETTER A..RAISED
1F16C..1F16F; UNASSIGNED  # <reserved>..<reserved>
1F170..1F1AC; DISALLOWED  # NEGATIVE SQUARED LATIN CAPITAL LETTER A..SQU
1F1AD..1F1E5; UNASSIGNED  # <reserved>..<reserved>
1F1E6..1F202; DISALLOWED  # REGIONAL INDICATOR SYMBOL LETTER A..SQUARED
1F203..1F20F; UNASSIGNED  # <reserved>..<reserved>
1F210..1F23B; DISALLOWED  # SQUARED CJK UNIFIED IDEOGRAPH-624B..SQUARED
1F23C..1F23F; UNASSIGNED  # <reserved>..<reserved>
1F240..1F248; DISALLOWED  # TORTOISE SHELL BRACKETED CJK UNIFIED IDEOGRA
1F249..1F24F; UNASSIGNED  # <reserved>..<reserved>
1F250..1F251; DISALLOWED  # CIRCLED IDEOGRAPH ADVANTAGE..CIRCLED IDEOGRA
1F252..1F25F; UNASSIGNED  # <reserved>..<reserved>
1F260..1F265; DISALLOWED  # ROUNDED SYMBOL FOR FU..ROUNDED SYMBOL FOR CA
1F266..1F2FF; UNASSIGNED  # <reserved>..<reserved>
1F300..1F6D4; DISALLOWED  # CYCLONE..PAGODA
1F6D5..1F6DF; UNASSIGNED  # <reserved>..<reserved>
1F6E0..1F6EC; DISALLOWED  # HAMMER AND WRENCH..AIRPLANE ARRIVING
1F6ED..1F6EF; UNASSIGNED  # <reserved>..<reserved>
1F6F0..1F6F9; DISALLOWED  # SATELLITE..SKATEBOARD
1F6FA..1F6FF; UNASSIGNED  # <reserved>..<reserved>
1F700..1F773; DISALLOWED  # ALCHEMICAL SYMBOL FOR QUINTESSENCE..ALCHEMIC
1F774..1F77F; UNASSIGNED  # <reserved>..<reserved>
1F780..1F7D8; DISALLOWED  # BLACK LEFT-POINTING ISOSCELES RIGHT TRIANGLE
1F7D9..1F7FF; UNASSIGNED  # <reserved>..<reserved>
1F800..1F80B; DISALLOWED  # LEFTWARDS ARROW WITH SMALL TRIANGLE ARROWHEA
1F80C..1F80F; UNASSIGNED  # <reserved>..<reserved>
1F810..1F847; DISALLOWED  # LEFTWARDS ARROW WITH SMALL EQUILATERAL ARROW
1F848..1F84F; UNASSIGNED  # <reserved>..<reserved>
1F850..1F859; DISALLOWED  # LEFTWARDS SANS-SERIF ARROW..UP DOWN SANS-SER
1F85A..1F85F; UNASSIGNED  # <reserved>..<reserved>
1F860..1F887; DISALLOWED  # WIDE-HEADED LEFTWARDS LIGHT BARB ARROW..WIDE
```

```
1F888..1F88F; UNASSIGNED  # <reserved>..<reserved>
1F890..1F8AD; DISALLOWED  # LEFTWARDS TRIANGLE ARROWHEAD..WHITE ARROW SH
1F8AE..1F8FF; UNASSIGNED  # <reserved>..<reserved>
1F900..1F90B; DISALLOWED  # CIRCLED CROSS FORMEE WITH FOUR DOTS..DOWNWAR
1F90C..1F90F; UNASSIGNED  # <reserved>..<reserved>
1F910..1F93E; DISALLOWED  # ZIPPER-MOUTH FACE..HANDBALL
1F93F       ; UNASSIGNED  # <reserved>
1F940..1F970; DISALLOWED  # WILTED FLOWER..SMILING FACE WITH SMILING EYE
1F971..1F972; UNASSIGNED  # <reserved>..<reserved>
1F973..1F976; DISALLOWED  # FACE WITH PARTY HORN AND PARTY HAT..FREEZING
1F977..1F979; UNASSIGNED  # <reserved>..<reserved>
1F97A       ; DISALLOWED  # FACE WITH PLEADING EYES
1F97B       ; UNASSIGNED  # <reserved>
1F97C..1F9A2; DISALLOWED  # LAB COAT..SWAN
1F9A3..1F9AF; UNASSIGNED  # <reserved>..<reserved>
1F9B0..1F9B9; DISALLOWED  # EMOJI COMPONENT RED HAIR..SUPERVILLAIN
1F9BA..1F9BF; UNASSIGNED  # <reserved>..<reserved>
1F9C0..1F9C2; DISALLOWED  # CHEESE WEDGE..SALT SHAKER
1F9C3..1F9CF; UNASSIGNED  # <reserved>..<reserved>
1F9D0..1F9FF; DISALLOWED  # FACE WITH MONOCLE..NAZAR AMULET
1FA00..1FA5F; UNASSIGNED  # <reserved>..<reserved>
1FA60..1FA6D; DISALLOWED  # XIANGQI RED GENERAL..XIANGQI BLACK SOLDIER
1FA6E..1FFFD; UNASSIGNED  # <reserved>..<reserved>
1FFFE..1FFFF; DISALLOWED  # <noncharacter>..<noncharacter>
20000..2A6D6; PVALID      # <CJK Ideograph Extension B>..<CJK Ideograph
2A6D7..2A6FF; UNASSIGNED  # <reserved>..<reserved>
2A700..2B734; PVALID      # <CJK Ideograph Extension C>..<CJK Ideograph
2B735..2B73F; UNASSIGNED  # <reserved>..<reserved>
2B740..2B81D; PVALID      # <CJK Ideograph Extension D>..<CJK Ideograph
2B81E..2B81F; UNASSIGNED  # <reserved>..<reserved>
2B820..2CEA1; PVALID      # <CJK Ideograph Extension E>..<CJK Ideograph
2CEA2..2CEAF; UNASSIGNED  # <reserved>..<reserved>
2CEB0..2EBE0; PVALID      # <CJK Ideograph Extension F>..<CJK Ideograph
2EBE1..2F7FF; UNASSIGNED  # <reserved>..<reserved>
2F800..2FA1D; DISALLOWED  # CJK COMPATIBILITY IDEOGRAPH-2F800..CJK COMPA
2FA1E..2FFFD; UNASSIGNED  # <reserved>..<reserved>
2FFFE..2FFFF; DISALLOWED  # <noncharacter>..<noncharacter>
30000..3FFFD; UNASSIGNED  # <reserved>..<reserved>
3FFFE..3FFFF; DISALLOWED  # <noncharacter>..<noncharacter>
40000..4FFFD; UNASSIGNED  # <reserved>..<reserved>
4FFFE..4FFFF; DISALLOWED  # <noncharacter>..<noncharacter>
50000..5FFFD; UNASSIGNED  # <reserved>..<reserved>
5FFFE..5FFFF; DISALLOWED  # <noncharacter>..<noncharacter>
60000..6FFFD; UNASSIGNED  # <reserved>..<reserved>
6FFFE..6FFFF; DISALLOWED  # <noncharacter>..<noncharacter>
70000..7FFFD; UNASSIGNED  # <reserved>..<reserved>
7FFFE..7FFFF; DISALLOWED  # <noncharacter>..<noncharacter>
80000..8FFFD; UNASSIGNED  # <reserved>..<reserved>
```

```
8FFFE..8FFFF; DISALLOWED  # <noncharacter>..<noncharacter>
90000..9FFFD; UNASSIGNED  # <reserved>..<reserved>
9FFFE..9FFFF; DISALLOWED  # <noncharacter>..<noncharacter>
A0000..AFFFD; UNASSIGNED  # <reserved>..<reserved>
AFFFE..AFFFF; DISALLOWED  # <noncharacter>..<noncharacter>
B0000..BFFFD; UNASSIGNED  # <reserved>..<reserved>
BFFFE..BFFFF; DISALLOWED  # <noncharacter>..<noncharacter>
C0000..CFFFD; UNASSIGNED  # <reserved>..<reserved>
CFFFE..CFFFF; DISALLOWED  # <noncharacter>..<noncharacter>
D0000..DFFFD; UNASSIGNED  # <reserved>..<reserved>
DFFFE..DFFFF; DISALLOWED  # <noncharacter>..<noncharacter>
E0000       ; UNASSIGNED  # <reserved>
E0001       ; DISALLOWED  # LANGUAGE TAG
E0002..E001F; UNASSIGNED  # <reserved>..<reserved>
E0020..E007F; DISALLOWED  # TAG SPACE..CANCEL TAG
E0080..E00FF; UNASSIGNED  # <reserved>..<reserved>
E0100..E01EF; DISALLOWED  # VARIATION SELECTOR-17..VARIATION SELECTOR-25
E01F0..EFFFD; UNASSIGNED  # <reserved>..<reserved>
EFFFE..10FFFF; DISALLOWED  # <noncharacter>..<noncharacter>
```

Author’s Address

Patrik Faltstrom
Netnod

Email: paf@netnod.se

IETF
Internet-Draft
Intended status: Standards Track
Expires: December 31, 2018

A. Freytag
ASMUS, Inc.
J. Klensin

A. Sullivan
Oracle Corp.
June 29, 2018

Those Troublesome Characters: A Registry of Unicode Code Points Needing Special Consideration When Used in Network Identifiers
draft-freytag-troublesome-characters-02

Abstract

Unicode's design goal is to be the universal character set for all applications. The goal entails the inclusion of very large numbers of characters. It is also focused on written language in general; special provisions have always been needed for identifiers. The sheer size of the repertoire increases the possibility of accidental or intentional use of characters that can cause confusion among users, particularly where linguistic context is ambiguous, unavailable, or impossible to determine. A registry of code points that can be sometimes especially problematic may be useful to guide system administrators in setting parameters for allowable code points or combinations in an identifier system, and to aid applications in creating security aids for users.

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 https://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 December 31, 2018.

Copyright Notice

Copyright (c) 2018 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Unicode code points and identifiers

Unicode [Unicode] is a coded character set that aims to support every writing system. Writing systems evolve over time and are sometimes influenced by one another. As a result, Unicode encodes many characters that, to a reader, appear to be the same thing; but that are encoded differently from one another. This sort of difference is usually not important in written texts, because competent readers and writers of a language are able to compensate for the selection of the "wrong" character when reading or writing. Finally, the goal of supporting every writing system also implies that Unicode is designed to properly represent written language; special provisions are needed for identifiers.

Identifiers that are used in a network or, especially, an Internet context present several special problems because of the above feature of Unicode:

[[CREF1: AF: This whole business of language context seems unconnected from the data we have in the registry: that data is about code points and sequences that look the same, and many examples are in the same language. For example the duplicated shapes for digit / letter pairs. In very few cases would knowing the language context make a difference. In some cases, if you knew the script (not for the label, but the code point) you might be able to distinguish two labels, but that is it. I think we should further rewrite this summary so it matches better with the what the proposed registry contains.]]

1. In many (perhaps most) uses of identifiers, they are neither constrained to words in a particular language, nor would it be possible to ascertain reliably the language context in which the identifier is being or will be used. In the case of an internationalized domain name, for instance, each label could in principle represent a new locus of control, because there could be a delegation there. A new locus of control means that the administrator of the resulting zone could speak, read, or intend a different language context than the one from the parent. Moreover, at least some domains (such as the root) have an Internet-wide context and therefore do not really have a language context as such. In any case, the language context is simply not available as part of a DNS lookup, so there is no way to make the DNS sensitive to this sort of issue. Even in the case of email local-parts, where a sender is likely to know at least one of the languages of the receiver, the language context that was in use at the time the identifier was created is often unknown.

2. Identifiers on the network are in general exact-match systems, because an ambiguous identifier is problematic. Sometimes, but not always, there are facilities for aliasing such that multiple identifiers can be put together as a single identity; the DNS, for example, does not have such an aliasing capability, because in the DNS all aliases are one-way pointers. Aliasing techniques are in any case just an extension of the exact-match approach, and do not work the way a competent human reader does when interpolating the "right" character upon seeing the "wrong" one.

3. Because there are many characters that may appear to be the same (or even, that are defined in such a way that they are all but guaranteed to be rendered by the same glyphs), it is fairly easy to create an identifier either by accident or on purpose that is likely to be confused with some other identifier even by competent readers and writers of a language. In some cases knowing the language context would be of no help to recognition, for example, in cases where a language uses the same shape for a letter as for one of the digits.

4. For some scripts their repertoire of shapes overlaps with one or more other scripts, so that there are cases where two strings look identical to each other, even though all the code points in the first string are of one script, and all the code points in the second string are of another script. In these cases, the strings cannot be distinguished by a reader, and the whole strings are confusable.

5. For some scripts, both users and rendering systems do not expect to encounter code points in arbitrary sequence. Most code points normally occur only in specific locations within a syllable. If random labels were permitted, some would not display as expected (including having some features misplaced or not displayed) while others would present recognition problems to users experienced with the script. Some devices may also not support arbitrary input.

Beyond these issues, human perception is easily tricked, so that entirely unrelated character sequences can become confusable -- for example "rn" being confused with "m". Humans read strings, not characters, and they will mostly see what they expect to see. Some additional discussion of the background can be found in Appendix A.

The remainder of this document discusses techniques that can be used to design the label generation rules for a particular zone so they ameliorate or avoid entirely some of the issues caused by the interaction between the Unicode Standard and identifiers. The

registry is intended to highlight code points that require such techniques.

2. Background and Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

A reader needs to be familiar with Unicode [Unicode], IDNA2008 [RFC5890] [RFC5891] [RFC5892] [RFC5893] [RFC5894], PRECIS (at least the framework, [RFC7564]), and conventions for discussion of internationalization in the IETF (see [RFC6365]).

3. Techniques already in place

In the IDNA mechanism for including Unicode code points [RFC5892], a code point is only included when it meets the needs of internationalizing domain names as explained in the IDNA framework [RFC5894]. For identifiers other than those specified by IDNA, the PRECIS framework [RFC7564] generalizes the same basic technique. In both cases, the overall approach is to assume that all characters are excluded, and then to include characters according to properties derived from the Unicode character properties. This general strategy cuts the enormous size of the Unicode database somewhat, avoiding including some characters that are necessarily unsuited for use as identifiers.

The mechanism of inclusion by derived property, while helpful, is insufficient to guarantee every included character is safe for use in identifiers. Some characters' properties lead them to be included even though they are not obviously good candidates. In other cases, individual characters are good for inclusion, but are problematic in combination. Finally, there are cases where characters (or sequences of characters) are not problematic by themselves, or if used in a mutually exclusive manner in the same identifier, but become problematic when their choice represents the only difference between otherwise identical identifiers. For some examples, see Appendix B.

Operators of systems that create identifiers (whether through a registry or through a peer-to-peer identifier negotiation system) need to make policies for characters they will permit. Operators of registries, for instance, can help by adopting good registration policies: "Users will benefit if registries only permit characters from scripts that are well-understood by the registry or its advisers."[RFC5894]

The difficulty for many operators, however, is that they do not have the writing system expertise to claim any character is "well-understood", and they do not really have the time to develop that expertise. Such operators should in fact not use or register such characters. Unfortunately, in many cases the operators are stewards of systems where the user population demands identifiers useful to them in their local languages. In other cases, operators may proceed without a proper understanding owing to financial or market share incentives. The risk for Internet identifiers in such cases is obviously that ill-understood and potentially exploitable gaps in registration policies will open.

To help mitigate such issues, this document proposes a registry of Unicode code points that are known to present special issues for network identifiers with the aim to guide protocol and operating decisions about whether to permit a given code point or sequence of code points. By necessity, any list or guidance can only reflect issues that are known and understood at the time of writing. By limiting itself largely to characters that are widely used to write languages in contemporary use, the registry will address the more critical needs, while simultanesously focusing on characters that are well understood and for which there may already be some implementation experience in IDNs.

By itself, such a registry will not completely protect against poor registration or use, but it may provide operational guidance necessary for people who are responsible for creating policies. It also obviates the need for everyone to repeat basic investigation into the behavior of Unicode characters. Instead, scarce expertise can be focused on ways to mitigate issues, perhaps caused by user requirements for a specific character.

Note that the registry defined herein does not address any of the issues created by whole-string confusables where each of the identifiers is of a different script. A common workaround, limiting a registry to identifiers of only a single script, would mitigate this issue. [[CREF2: AF: we should evaluate that; cross-script variants that are homoglyphs have now been collected across modern scripts as part of the root zone LGR and are easily captured in a registry.]]

For some of the code points (or code point sequences) listed as presenting issues for identifiers, it may be most expeditious to simply not include them, even though they are valid according to the protocol. Sometimes, one of a pair of identical code points (or code point sequences) may be deemed preferable over the other for practical reasons.

However, simply leaving out any code point listed in this registry would render a registry of doubtful value for many scripts. It is not always necessary or desirable to exclude characters. Sometimes, it is merely necessary to ensure that for two otherwise identical identifiers, only one of a set of mutually exclusive code points (or sequences of code points) is used, while preventing the later registration of the label containing the other one in order to avoid ambiguity. This way the operator does not need to impose a choice.

In cases where two or more variants of such an identifier mean the same thing to the native reader, an operator may decide to allow all of the variant labels to be registered simultaneously, but only to the same entity (and with proper safeguards that limit the multiplicity of such allocatable variant labels).

The implementation of this strategy would be via the variant mechanism described in [RFC7940] and [RFC8228] which allows mechanical processing of mutual exclusion and /or bundling of identifiers respectively.

This specification defines a registry of code points and sequences that have been identified as requiring special attention when they are to be used in identifiers. An administrator who does not have the time or inclination to develop the requisite policies might contemplate simply not to permit these code points at all.

However, for some scripts the remaining subset might not be usable in a meaningful way. Identifiers in these scripts cannot be safely implemented without understanding the issues involved. Further note that many code points listed here are problematic only in their relationship to other code points and that as long as these issues are adequately addressed, for example using the variant mechanism, they do not need to be excluded. [[CREF3: AF: the above needs more editing, it’s a bit repetitive.]]

4. A registry of code points requiring special attention

4.1. Description

The registry contains four fields. [[CREF4: AF If we are limited to the "texttable" format, we are limited to three columns, there’s no way we can fit more than that into the RFC plain text format and remain legible. If we want more columns, then we need to use some other data format, including PDF (which would allow us to show the images for the code points).]]

1. The first field, called "Code Point(s)", is a code point or sequence of code points. Sequences in this and other fields are

listed as space separated code point values. For completeness, full code point sequences are listed, even if some of their constituents are "Not recommended". A code point value is a series of 4-6 uppercase hexadecimal digits, as defined in [Unicode].

2. The second field, "Related CP", contains zero or more cross references to related code points or sequences. Cross references consist of single code points or sequences. Multiple cross references are separated by a comma.

3. The third field, called "References", contains one or more references to documents describing the code point and the reason why it presents an issue. References are cited by numeric values, each in square brackets; multiple references are separated by space.

4. The last field, "Comment", is a free form text field that briefly describes the issue; it also The comment field starts with a category, separated by a colon, to allow quick identification of similar cases

The following are the defined category values:

Not Recommended While the code point (or sequence) is not DISALLOWED, there is emerging consensus in the community that it is not recommended for identifiers, or it is considered as such in the Unicode Standard. This includes, but is not limited to code points that are formally deprecated in the Unicode standard, as well as code points or sequences listed in the standard as "Do not use" or not preferred or similar. Code points not in active use, obsolete code points, or those intended for specialist use may also be listed under this category. Details are given in the explanation and references.

Identical The code point (or sequence) is normally identical in appearance to another code point (or sequence); or may be identical in some contexts. If the related CP is listed as "PREFERRED", it is recommended that this code point (or sequence) be excluded; in the case of a sequence, it may be appropriate to exclude, the constituent combining marks (after first consulting the details given in the listing for the marks). Otherwise, it is recommended to make the two identical code points or sequences mutually exclusive by treating them as variants. Details are given in the explanation and references.

Restricted Context The code point is problematic in relation to some other code points in the same label. For example, it should be

used only after some code points or not adjacent to certain other code points. Further details are given in the explanation and references. This is a common case for certain combining marks or other code points in so-called "complex" scripts. These scripts generally require a coordinated set of context rules; in those cases the registry would not list any specific context rules, but to point to documentation of existing Label Generation Rulesets implementing a coherent set of rules as examples. Code points with IDNA2008 property of CONTEXTJ or CONTEXTO are not listed, as long as the given context rules mitigate any concerns.

Preferred The code point is preferred to some other code point given in the cross reference (with the other code point normally "IDENTICAL" or "NOT RECOMMENDED"). In some cases this represents a preference for a code point (or sequence) that is a basic constituent in some alphabet over a code point (or sequence) that is rare or has specialized use. In some cases the preference may be formally specified or otherwise represent established community consensus. Details are given in the explanation and references.

Other All cases that do not fit one of the other categories. Details are given in the explanation and references.

If a character appears in the registry, that does not automatically mean that it is a bad candidate for use in identifiers generally. Absent a well-defined and verifiable policy, however, such a code point or sequence might well be treated with suspicion by users and by tools.

For code points tagged as being "identical" to or "indistinguishable" from other code points, it may be that one is preferred over the other, but it may also be that implementing a scheme for mutual exclusion of any resulting identical labels is the best solution, such as assigning them "blocked" variants according to [RFC7940] and [RFC8228].

Where characters are confusable with a combining sequence, only the combining sequence is listed; suggested mitigation may consist of disallowing either the specific combining sequence or disallowing the combining marks involved. It is usually inappropriate to exclude any of the basic letters involved, as they are generally members of the standard alphabet for one or more languages.

The registry and this document are to be understood as guidance for the purpose of developing operational policies that are used for protocols under normal administrative scope. For instance, zone operators that support IDNA are expected to create policies governing the code points that they will permit (see [RFC5894] and

[I-D.rfc5891bis]). The registry herein defined is intended to highlight particularly troublesome code points or code point sequences for the benefit of administrators creating such policies. It is also intended to highlight characters that may create identifier ambiguities and thereby create security vulnerabilities. However, by itself it is no substitute for such policies.

The registry is by necessity limited to code points for which adequate information is available; by and large this means code points used in connection with modern languages or writing systems, except that specialized extensions to modern scripts may be indicated, if their use would fall into any of the categories defined. Historic scripts, and any modern scripts not represented in the registry can be assumed to not be well-understood; operators are cautioned to locate other sources of information and to develop the necessary policies before deploying such scripts.

4.2. Maintenance

The registry is updated by Expert Review using an open process. From time to time, additional code points may be added to the Unicode standard, or further information may be discovered related to code points, to existing code points or those already listed here. The Unicode Standard may recommend against using a code point for all or some purposes. Or a script community may have gained more experience in deploying IDNs for that script and may create or update recommendations as to best policy.

4.3. Scope

Code points that are DISALLOWED in IDNA 2008 are not eligible to be listed. Code points that are CONTEXTJ or CONTEXTO are not included here unless there are documented concerns that are not mitigated by the existing IDNA context rules. The focus is on scripts that are significant for identifiers; code points from scripts that are historic or otherwise of limited use have generally not been considered - however exceptions may exist where authoritative information is readily available. Code points and code point sequences included are those that need special policies (including, but not limited to policies of exclusion).

New code points of sequences are listed whenever information becomes available that identifies a specific issue that requires attention in crafting a policy for the use of that code point or sequence in network identifiers. Likewise cross references, categories, explanations and references cited may be updated.

The contents of the registry generally does not represent original research but a collection of issues documented elsewhere, with appropriate references cited. An exception might be cases that are in clear analogy to existing entries, but not explicitly covered by existing references, for example, because the code point in question was recently added to Unicode.

If a particular language or script community reaches an apparent consensus that some code point is problematic, or that of two identical code points or sequences one should be preferred over the other, such recommendations, if known, should be documented in this registry.

In addition, if the Unicode Standard designates a code point as formally "deprecated" or less formally as "do not use", or identifies code points that are "intentionally identical", this is also something that should be reflected in the registry. Another source of potential information might be existing registry policies or recommended policies, particularly where it is apparent that they represent a careful analysis of the issue or a wider consensus, or both.

Proposed additions to the registry are to be shared on a mailing list to allow for broader comment and vetting.

If there is a disagreement about the existence of an issue or its severity, it is preferable to document both the issue and the different evaluations of it. In all cases, the information and documentation presented must allow a user to fully evaluate the status of any entry in the registry.

There is no requirement for the registry to form a stable body of data to which any future document would have to be backward compatible in any way. If new information emerges, additional code points may be considered problematic, or they may need to be reclassified. In case of significant changes, the explanation should note the nature of the change and cite a reference to document the basis for it.

5. Registry initial contents

5.1. Overview

IDNA 2008 uses an inclusion process based on Unicode properties to define which code points are PVALID, but also recognizes that some code points require a context rule (CONTEXTJ, CONTEXTO).

A number of code points which are PVALID in [RFC5892] may require additional attention in the design of label generations rules. In some cases, the issue is not necessarily with an individual code point, but with a code point sequence. In the following, "code point" and "code point sequence" are used synonymously unless explicitly called out. The fact that a code point require such attention does not affect its status under IDNA 2008.

The following describes a number of conditions that pose problems for network identifiers and common strategies for mitigating them.

5.2. Interchangeable Code Points

At times two code points or code point sequences are considered by all users (or a significant fraction) as equivalent to a degree that they accept one of them as substitute for another. This has obvious implications for the unambiguous recognition of identifiers. This document lists the code points and sequences affected (except for certain generic classes too numerous to list here). Note that one of the two may be preferred over the other, in which case the non-preferred one may be excluded or folded away. But in many cases either one is equally preferred. Mitigation techniques for such cases are discussed below.

Homoglyphs Homoglyphs are code points that have identical appearance, or are so close in appearance that they are indistinguishable if not presented side-by-side. Whenever two labels differ only by code points that are homoglyphs of each other and occur in the same position, users cannot distinguish the labels from each other or tell which label is intended, even though the underlying code points are different. Users will substitute one label for another.

Code points that are merely similar in appearance, including strongly similar code points, or code points that are difficult to distinguish (such as certain diacritical marks) are not considered here; handling such similarities often requires case by case judgment.

Instead, this document considers these types of code points that can be fully substituted for one another:

1. code points that, by design or derivation, are identical to each other;

2. code points that assume the same shape in some context, e.g. at the end of a label;

3. code points of a striking similarity based on derivation or common origin;

4. and code points that are otherwise indistinguishable from one another unless placed side by side.

Cross-script Homoglyphs A number of code points are homoglyphs of code points in another script (cross-script homoglyphs). Cross-script homoglyphs are a concern for any zone that supports labels from more than one script, even if each label is required to be in a single script. Note that some writing systems ordinarily use a combination of scripts (such as the use of Han, Hiragana and Katakana for Japanese). For many writing systems, an admixture of Latin letters is not uncommon, for example in brand or product names. If not handled carefully, this can prove problematic for identifiers.

Homophones As discussed in [202], the Amharic language treats many code points from the Ethiopic script as sound-alikes (homophones). In writing, these are freely substituted, users do not recognize some spelling as more correct. A conservative approach would treat these as mutually exclusive; the alternative, to make all variants available to the same applicant is appears not feasible due to the high number of such variants per label.

Semantic Variants The Chinese writing system, shared among several geographically distributed user communities, has many instances of code points that represent the same semantic. Even though they are visually distinct, they can be substituted for one another; typically these correspond to the simplified and traditional forms of Chinese characters. See [RFC4713] for details.

5.3. Excludable Code Points

Code points that are not substitutable but troublesome for other reasons are candidates for exclusion from a zone’s repertoire. For each such code point, the comment field briefly describes why it should be excluded or considered troublesome. There is no identified mitigation strategy that can be recommended for general usage: unless careful study indicates that a code point with this status is exceptionally acceptable for a particular zone, after all, it should normally be excluded from the repertoire. These reasons are varied.

Deprecated Code Points Deprecated code points are those that [Unicode] recommends not to use for any purpose. They should be excluded from identifiers; there is no mitigation. In addition, Unicode recommends against the use of some sequences and code

points for any purpose, but without formal deprecation. These should likewise be excluded from identifiers.

Non-preferred or other Troublesome Code Point This category includes all code points that are troublesome for other reasons; they include code points that represent non-preferred variations; or code points that not meant to be used in a combining sequence for letter; or code points that may be indistinguishable from a punctuation mark or other DISALLOWED code point. For each such code point, the comment field briefly describes why it should be excluded or considered troublesome.

Obsolete or not in Active Use Many code points across scripts that are otherwise in modern use represent additions for use in obsolete orthographies and writing systems, that is for writing languages that are extinct or not longer written in that script. Some have been researched and no evidence of active use could be found. These code points are not recommended for use in identifiers and should be excluded. Except for specialists, users are unlikely to recognize them, or find them of use in constructing mnemonic strings for identifiers. In addition, they often have not been sufficiently analyzed as to whether they represent other issues for identifiers. That makes their use risky. Obsolete, rare and code points otherwise not in active are generally not listed here. The reader can find a list of code points with high probability of being in active use in [MSR].

5.4. Combining Marks

Non Normalizable Sequences Certain combining marks are part of non-normalizable sequences. Normally, when a combining sequence is an alternate encoding to a composite code point, normalization can be used to select a preferred representation. For IDNA 2008, which uses NFC to normalize, this means the composite code point. However, some combining marks are not considered identical to the same mark when graphically part of a composite character. Sequences with these marks may look more or less like some composite code point, but they are considered different, and therefore not normalized. For identifiers, the best recommendation is to exclude those combining marks.

Combining marks that are also part of precomposed letters Many combining marks are part of canonical decompositions. For identifiers that are normalized to the composed forms using NFC (as required by IDNA 2008), these combining marks usually are not needed on their own, that is as separate element of a combining sequence after normalization. (The vast majority of letters using these marks have been encoded as precomposed characters). It is

strongly recommended to exclude these combining marks on their own, but, as needed for a specific language, to enumerate the needed sequences. (One notable example is Vietnamese which, after normalization to NFC uses a mixture of precomposed code points and combining marks). [TBD]The most common generic combining marks affected have been entered in the registry as excluded.

Non-spacing combining marks These marks are typically accents, diacritics and the like. They pose an additional problem: if they are allowed to occur twice in a row, some rendering systems will "overprint" them, in effect making them indistinguishable from single marks. This problem can be avoided by allowing only enumerated sequences, or alternatively by a context rule.

Ambiguous Rendering There are other ways in which certain code points and sequences representing particular combinations of code points may suffer from unreliable rendering, because rendering engines normally do not expect to encounter them. While Unicode allows the use of combining marks, in principle, in combination with any base character, in practice this can lead to unrecognizable labels, or labels that are not reliably distinct. This situation mostly affects the so-called complex scripts.

Combining marks in complex scripts In some scripts, there are no precomposed sequences. Usually, these scripts are "complex" scripts, that require context rules for many classes of code points. For these scripts, context rules (see [RFC7940]) should be used to limit non-spacing marks to acceptable contexts. For an example of such rules see [204], [206].

Soft Dotted and Dotless Letters Unicode code points with the Soft_Dotted property encode letter that lose their dot if followed by a diacritical mark above. (See [UCD]) If the following mark is a COMBINING DOT ABOVE, the combination is indistinguishable from the letter by itself. This can be mitigated by limiting or excluding the code point for DOT ABOVE. A soft dotted code point followed by any other diacritical mark above will look identical to the corresponding dotless letter with diacritical mark above. All combinations of dotless letters followed by diacritical marks should be excluded. (This can be done with a context rule, see [RFC7940]).

5.5. Mitigation

Thiere are several techniques that can be used to help to mitigate confusion. The focus in the following is on issues addressable by protocol or registry policy. However, user agents might implement

additional mitigation approaches, such as always using a font designed to distinguish among different characters.

5.5.1. Mitigation Strategies

Exclusion The primary mitigation technique is to reduce the problem space: operators should only ever use the smallest repertoire of code points possible for their environment. So, for example, if there is a code point that is sometimes used but is perhaps a little obscure, it is better to leave it out. Users are unlikely to be familiar with many code points added to Unicode for the representation of historical forms of writing a script, or for highly specialized purposes. That unfamiliarity may present challenges to correct identification or keyboard entry, making the code point less usable. In addition, their use may present other problems not appreciated by anyone not familiar with them.

For these reasons, code points used only in a language with which the administrator is not familiar should probably be excluded. The same applies to code points used in specialized contexts, such as those only found in historic or sacred documents, or only used for phonetic transcription or poetry.

By reducing the repertoire to a well-understood essential subset it is often possible to eliminate some possible instances of confusion. For example, in the Arabic script, combining marks are generally used for optional or specialized aspects of the writing system. At the same time, many combining sequences are confusable with basic letters of the script. Because of this, excluding all Arabic combining mark would greatly reduce confusability without significantly affecting usability of the script for identifiers.

Preferred code points Sometimes, each of these code points will be used by a different user community; or one of the code points is not in wide use, for example because it is intended for special purposes like phonetic annotation or transliteration. In such cases, the one not needed for a given zone could be excluded.

In other cases, zones may be shared by a wider community, making it unattractive or impossible to institute a preference. A common method of mitigating issues from such homoglyphs is to make two labels that differ only by using a different homoglyph mutually exclusive. This can be done by making the homoglyphs code point variants, usually of type "blocked". See [RFC8228].

In some cases, while two code points may be homoglyphs, one of them can be identified as the preferred alternative to encode the intended character. In these cases, one of the code points has

been identified as "preferred", while the other has been identified as "troublesome"; or "excluded". In all other cases, no such preference exists in the general usage; a conservative mitigation might be to define the alternatives as blocked variants. However, the users of a given zone might have a specific preference, in which case one of the alternatives could be excluded instead.

For convenience in presentation, this document presents pairs or sets of homoglyphs as mutually exclusive variants of type "homoglyph". Other ways of handling these code points are possible. While one might implement such a variant relation in many cases as one label blocking another, in some cases allowing both to be registered to the same applicant may be appropriate. Finally, in some case eliminating one or both code points from the repertoire may be a feasible alternative to establishing a variant relation.

Script limitation For homoglyphs, a large number of cases (but not all of them) turn out to be in different scripts. As a result, it is usually a good idea to adopt the operational convention that identifiers for a protocol should always be in a single script.

This mitigation strategy has limits. First, even if any given identifier is only in a single script, it may co-exist with identifiers from other scripts. Sometimes the repertoire used in operation allows multiple scripts that create whole string confusables -- strings made up entirely of homoglyphs of another string in a different script (such as can be found between Cyrillic and Latin, for example). In such cases, mitigation must turn to other means of preventing the registration of mutually confusable string, for example by In that case, a robust mechanism for mutual exclusion of confusable identifiers must exist, ensuring that the registration of one of them (whichever comes first) blocks the later registration of the other.

Second, some writing systems use a combination of scripts and for commercial names in many scripts, admixture of Latin letters is common. Allowing limited script mixing may be an essential requirement in some cases.

Lastly, identifiers are not always under the operational control of a single authority (such as in the case of DNS, where the system is under distributed control so that different parts of the hierarchy can have different operational rules).

In the case of IDNA, some client programs restrict display of U-labels to top-level domains known to have policies about single-script labels.

Exact homoglyphs No policy or convention, other than ensuring mutual exclusion, will do anything to help mititgate confusion for strict homoglyphs of each other in the same script (see Appendix B for some example cases.)

Beyond the issue of mutual confusability, some combining sequences in particular can give rise to other difficulties in recognition - usually because client systems will not reliably and correctly display them. One particular case concerns sequences of more than one instance of the same non-spacing combining mark such as the repetition of an accent or diacritic. These are often rendered indistinguishably from single instances of the same mark. Operators should prohibit such repetition, particularly, as there are no known cases where they would be required in ordinary writing. Note that this prohibition would also apply to a non-spacing mark following a pre-composed code point containing the same diacritic. A more general mitigation technique would be to limit nonspacing marks to known combinations which can be enumerated. Where that is not possible for some scripts, some other context restrictions can usually be applied.

There are some writing systems where characters do not normally occur in arbitrary locations in the context of each syllable. Neither users nor rendering systems for such scripts are adept at handling arbitrary sequences of such characters. While some latitude beyond strict spelling rules may be accommodated, policies that enforce a minimal set of structural rules are required to ensure that users can identify the identifier and systems can render them predictably.

5.5.2. Limits of Mitigation

As noted in Section 1, it is not possible to solve all the problems with identifier systems, particularly when human factors are taken into account. In addition, each of the mitigation approaches has its own limits of the type of problems that can be addressed, whether it is by exclusion of specific code points; requiring or prohibiting contexts for certain code points; restriction to a single script per label; or mutual exclusion of labels differing only by code points identical or otherwise confusably equivalent to other code points. Additional policies may be needed to prevent registration of labels that are problematic or confusable for other reasons.

There are a number of issues in implementing and presenting identifiers to the user which are not specific to individually identifiable code points (or sequences). For example, fonts can vary widely in whether they make or do not make a distinction in appearance of characters; relying on the native reader to get the intended meaning from context. It is up to user agents to make sure to select fonts that render each code point as distinct as possible.

When new code points are assigned in Unicode, systems, keyboards, fonts and rendering engines may all be updated unevenly, with considerable delays. During a possibly lengthy transition period, this will lead to inconsistent user experience or inability to distinguish certain labels. Even if unsupported labels are presented as A-labels, users may not reliably identify them, because they appear as essentially random sequences of letters and digits.

5.6. Notes

In the explanation the character names have been abbreviated. The following list shows sample entries for the proposed registry. It is non-normative, and only included for illustrative purposes. Also see the examples below (Appendix B).

6. Table of Code Points

```
-----------------------------------------------------------------
  Code Point: 01C0
  Related CP:
  References: [120] [155]
  Comment:    Not Recommended: Indistinguishable from a
              punctuation character that is not PVALID
-----------------------------------------------------------------
  Code Point: 01C1
  Related CP:
  References: [120] [155]
  Comment:    Not Recommended: Indistinguishable from a
              punctuation character that is not PVALID
-----------------------------------------------------------------
  Code Point: 01C2
  Related CP:
  References: [120]
  Comment:    Not Recommended: Indistinguishable from a
              punctuation character that is not PVALID
-----------------------------------------------------------------
  Code Point: 01C3
  Related CP:
  References: [120] [150]
  Comment:    Not Recommended: Indistinguishable from a
```

```
              punctuation character that is not PVALID
   ---------------------------------------------------------------
     Code Point: 01DD
     Related CP: 0259
     References: [150]
     Comment:    Identical: Identical in appearance to U+0259
   ---------------------------------------------------------------
     Code Point: 0259
     Related CP: 01DD
     References: [150]
     Comment:    Identical: Identical in appearance to U+01DD
   ---------------------------------------------------------------
     Code Point: 0131
     Related CP:
     References: [100]
     Comment:    Restricted Context: If followed by any combining
                 mark above, renders the same way as U+0069 in any
                 good font. Should be restricted to where it is not
                 followed by a combining mark above
   ---------------------------------------------------------------
     Code Point: 0237
     Related CP:
     References: [115]
     Comment:    Not Recommended: If followed by any combining mark
                 above, renders the same way as U+006A in any good
                 font. As its use is limited, it is best excluded.
   ---------------------------------------------------------------
     Code Point: 025F
     Related CP:
     References: [115]
     Comment:    Not Recommended: If followed by any combining mark
                 above, renders the same way as U+0249 in any good
                 font. As its use is limited, it is best excluded.
   ---------------------------------------------------------------
     Code Point: 02A3
     Related CP: 0064 007A
     References: [115]
     Comment:    Not Recommended: Looks like small LETTER D plus
                 LETTER Z, except for slight kerning; in limited
                 use.
   ---------------------------------------------------------------
     Code Point: 02A6
     Related CP: 0074 0073
     References: [115]
     Comment:    Not Recommended: Looks like small LETTER T plus
                 LETTER S, except for slight kerning; in limited
                 use.
   ---------------------------------------------------------------
```

```
  Code Point: 02A7
  Related CP: 0074 0283
  References: [115]
  Comment:    Not Recommended: Looks like small LETTER T plus
              LETTER ESH, except for slight kerning; in limited
              use.
---------------------------------------------------------------
  Code Point: 02AA
  Related CP: 006C 0073
  References: [115]
  Comment:    Not Recommended: Looks like small LETTER L plus
              LETTER S, except for slight kerning; in limited
              use.
---------------------------------------------------------------
  Code Point: 02AB
  Related CP: 006C 007A
  References: [115]
  Comment:    Not Recommended: Looks like small LETTER L plus
              LETTER Z, except for slight kerning; in limited
              use.
---------------------------------------------------------------
  Code Point: 02B9
  Related CP:
  References: [120]
  Comment:    Not Recommended: Indistinguishable from a
              punctuation character that is not PVALID
---------------------------------------------------------------
  Code Point: 02BA
  Related CP:
  References: [120]
  Comment:    Not Recommended: Indistinguishable from a
              punctuation character that is not PVALID
---------------------------------------------------------------
  Code Point: 02BB
  Related CP:
  References: [120]
  Comment:    Not Recommended: Indistinguishable from a
              punctuation character that is not PVALID
---------------------------------------------------------------
  Code Point: 02BC
  Related CP:
  References: [6912]
  Comment:    Not Recommended: Indistinguishable from a
              punctuation character (U+2019), which is not
              PVALID
---------------------------------------------------------------
  Code Point: 02BD
  Related CP:
```

```
  References: [120]
  Comment:    Not Recommended: Indistinguishable from
              punctuation character that is not PVALID
---------------------------------------------------------------
  Code Point: 02BE
  Related CP:
  References: [120]
  Comment:    Not Recommended: Indistinguishable from
              punctuation character that is not PVALID
---------------------------------------------------------------
  Code Point: 02BF
  Related CP:
  References: [120]
  Comment:    Not Recommended: Indistinguishable from
              punctuation character that is not PVALID
---------------------------------------------------------------
  Code Point: 02C0
  Related CP:
  References: [120]
  Comment:    Not Recommended: Indistinguishable from
              punctuation character that is not PVALID
---------------------------------------------------------------
  Code Point: 02C1
  Related CP:
  References: [120]
  Comment:    Not Recommended: Indistinguishable from
              punctuation character that is not PVALID
---------------------------------------------------------------
  Code Point: 02C6
  Related CP:
  References: [120]
  Comment:    Not Recommended: Indistinguishable from
              punctuation character that is not PVALID
---------------------------------------------------------------
  Code Point: 02C7
  Related CP:
  References: [120]
  Comment:    Not Recommended: Indistinguishable from
              punctuation character that is not PVALID
---------------------------------------------------------------
  Code Point: 02C8
  Related CP:
  References: [120]
  Comment:    Not Recommended: Indistinguishable from
              punctuation character that is not PVALID
---------------------------------------------------------------
  Code Point: 02C9
  Related CP:
```

```
  References: [120]
  Comment:    Not Recommended: Indistinguishable from
              punctuation character that is not PVALID
---------------------------------------------------------------
  Code Point: 02CA
  Related CP:
  References: [120]
  Comment:    Not Recommended: Indistinguishable from
              punctuation character that is not PVALID
---------------------------------------------------------------
  Code Point: 02CB
  Related CP:
  References: [120]
  Comment:    Not Recommended: Indistinguishable from
              punctuation character that is not PVALID
---------------------------------------------------------------
  Code Point: 0300
  Related CP:
  References: [100]
  Comment:    Not Recommended: Not recommended other than as
              part of enumerated sequences
---------------------------------------------------------------
  Code Point: 0301
  Related CP:
  References: [100]
  Comment:    Not Recommended: Not recommended other than as
              part of enumerated sequences
---------------------------------------------------------------
  Code Point: 0302
  Related CP:
  References: [100]
  Comment:    Not Recommended: Not recommended other than as
              part of enumerated sequences
---------------------------------------------------------------
  Code Point: 0303
  Related CP:
  References: [100]
  Comment:    Not Recommended: Not recommended other than as
              part of enumerated sequences
---------------------------------------------------------------
  Code Point: 0304
  Related CP:
  References: [100]
  Comment:    Not Recommended: Not recommended other than as
              part of enumerated sequences
---------------------------------------------------------------
  Code Point: 0306
  Related CP:
```

```
  References: [100]
  Comment:    Not Recommended: Not recommended other than as
              part of enumerated sequences
---------------------------------------------------------------
  Code Point: 0307
  Related CP:
  References: [115]
  Comment:    Restricted Context: By definition, LATIN SMALL
              LETTER I plus combining DOT ABOVE renders exactly
              the same as LATIN SMALL LETTER I by itself and
              does so in practice for any good font. The same is
              true for all Unicode characters with the
              soft_dotted property; they lose their dot if
              followed by a combining mark. DOT ABOVE should be
              excluded, or restricted to  contexts where it does
              not follow a soft_dotted letter.
---------------------------------------------------------------
  Code Point: 0308
  Related CP:
  References: [100]
  Comment:    Not Recommended: Not recommended other than as
              part of enumerated sequences
---------------------------------------------------------------
  Code Point: 0624
  Related CP: 0648
  References: [201]
  Comment:    Identical: Identical in appearance in some
              positional form and/or not reliably distinguished
              because of small size of distinguishing features
---------------------------------------------------------------
  Code Point: 0625
  Related CP: 0622, 0623, 0627, 0672
  References: [201]
  Comment:    Identical: Identical in appearance in some
              positional form and/or not reliably distinguished
              because of small size of distinguishing features
---------------------------------------------------------------
  Code Point: 0626
  Related CP: 0649, 064A, 067B, 06CC, 06CD, 06D0, 06D2
  References: [201]
  Comment:    Identical: Identical in appearance in some
              positional form and/or not reliably distinguished
              because of small size of distinguishing features
---------------------------------------------------------------
  Code Point: 0627
  Related CP: 0622, 0623, 0625, 0672
  References: [201]
  Comment:    Identical: Identical in appearance in some
```

positional form and/or not reliably distinguished because of small size of distinguishing features

---

Code Point: 064B
Related CP:
References: [5564]
Comment: Not Recommended: Not to be used in zone files for the Arabic language, per RFC 5564

Code Point: 064C
Related CP:
References: [5564]
Comment: Not Recommended: Not to be used in zone files for the Arabic language, per RFC 5564

Code Point: 065C
Related CP:
References: [300]
Comment: Not Recommended: Part of homoglyph sequence(s) not covered by normalization.

Code Point: 0660
Related CP: 06F0
References: [110]
Comment: Identical: Identical in appearance and meaning to EXTENDED ARABIC-INDIC DIGIT ZERO

Code Point: 0661
Related CP: 06F1
References: [110]
Comment: Identical: Identical in appearance and meaning to EXTENDED ARABIC-INDIC DIGIT ONE

Code Point: 077F
Related CP:
References: [115]
Comment: Not Recommended: Obsolote (archaic)

Code Point: 08AA
Related CP:
References: [201]
Comment: Not Recommended: No evidence of active use found; not recommended

---

Code Point: 0A72 0A3F
Related CP: 0A07
References: [401]
Comment: Not Recommended: Do not use for U+0A07

```
---------------------------------------------------------------
  Code Point: 0A72 0A40
  Related CP: 0A08
  References: [401]
  Comment:    Not Recommended: Do not use for U+0A08
---------------------------------------------------------------
  Code Point: 0E3A
  Related CP:
  References: [206]
  Comment:    Other issue: Renders unreliably, or not at all, if
              adjacent to any Thai vowel below. This may be
              prevented by a context rule
---------------------------------------------------------------
  Code Point: 0E41
  Related CP:
  References: [206]
  Comment:    Restricted Context: Digraph of U+0E40 SARA E
              U+0E40 SARA E. Normally handled by disallowing the
              sequence via a context rule
---------------------------------------------------------------
  Code Point: 0E45
  Related CP:
  References: [206]
  Comment:    Restricted Context: Only occurs after two special
              Thai vowels,U+0E24 RU and U+0E26 LU. Is also
              potentially confused with U+0E32 SARA I. Both
              issues can be addressed by defining a context
              rule. Alternatively the context may be spelled out
              by enumerating the two sequences and excluding
              U+0E45 if occurring by itself.
---------------------------------------------------------------
  Code Point: 0E4E
  Related CP:
  References: [206]
  Comment:    Not Recommended: Rarely used in modern Thai; it is
              more commonly replaced with U+0E3A (PHINTHU).
              Excluding it avoids issues with confusing it with
              another diacritic U+0E4C (THANTHAKHAT). Both are
              rendered atop a syllable and hard to distinguish
              at small sizes.
---------------------------------------------------------------
  Code Point: 12A5
  Related CP: 12D5
  References: [100] [202]
  Comment:    Interchangeable: U+12A5 and U+12D5 are used
              interchangeably in Amharic
---------------------------------------------------------------
  Code Point: 12A6
```

```
  Related CP: 12D6
  References: [100] [202]
  Comment:    Interchangeable: U+12A6 and U+12D6 are used
              interchangeably in Amharic
---------------------------------------------------------------
  Code Point: 17D2 178A
  Related CP: 17D2 178F
  References: [204]
  Comment:    Identical: When preceded by U+17D2, U+178A and
              U+178F are indistinguishable
---------------------------------------------------------------
  Code Point: 17D2 178F
  Related CP: 17D2 178A
  References: [204]
  Comment:    Identical: When preceded by U+17D2, U+178A and
              U+178F are indistinguishable
---------------------------------------------------------------
```

6.1. References for Registry

[99] The Unicode Consortium, "The Unicode Standard", (latest version) http:www.unicode.org/versions/latest (Multiple, or latest version)

[100] Integration Panel, "Maximal Starting Repertoire (MSR-2)", April 2015, https://www.icann.org/en/system/files/files/msr-2-overview-14apr15-en.pdf (Code points included in MSR-2 as potentially appropriate for the root zone)

[115] Integration Panel, "Maximal Starting Repertoire (MSR-2)", April 2015, https://www.icann.org/en/system/files/files/msr-2-overview-14apr15-en.pdf (Code points excluded from MSR-2 as inappropriate for the root zone)

[120] Integration Panel, "Maximal Starting Repertoire (MSR-2)", April 2015, https://www.icann.org/en/system/files/files/msr-2-overview-14apr15-en.pdf (Code points considered problematic by MSR-2)

[150] The Unicode Consortium, "Intentional.txt", Version 10.0.0, http://www.unicode.org/Public/security/10.0.0/intentional.txt (Code points considered identical by intention)

[155] "Proposal to Update Identical.txt", L2 17/301 (and revisions) http://www.unicode.org/L2/L2017/17301-update-intentional.pdf (Code points considered identical by intention)

[201] TF-AIDN, "Proposal for Arabic Script Root Zone LGR", 18 November 2015 https://www.icann.org/en/system/files/files/arabic-lgr-proposal-18nov15-en.pdf (In-script variants and code points excluded)

[202] Ethiopic Generation Panel, "Proposal for Ethiopic Script Root Zone LGR", May 17, 2017, https://www.icann.org/en/system/files/files/proposal-ethiopic-lgr-17may17-en.pdf ()

[204] Khmer Generation Panel, "Proposal for Khmer Script Root Zone Label Generation Rules (LGR)", August 15, 2016, https://www.icann.org/en/system/files/files/proposal-khmer-lgr-15aug16-en.pdf ()

[206] Thai Generation Panel, "Proposal for the Thai Script Root Zone LGR", May 25, 2017 https://www.icann.org/en/system/files/files/proposal-thai-lgr-25may17-en.pdf ()

[300] Internationalized Domain Names Variant Issues Project: Arabic Case Study Team Issues Report, ICANN, October 7, 2011 https://archive.icann.org/en/topics/new-gtlds/arabic-vip-issues-report-07oct11-en.pdf (In-script variants and code points excluded)

[401] Table 12-14 in Chapter 12 "South and Central Asia-I", ,"The Unicode Standard", Version 10.0, https://www.unicode.org/versions/Unicode10.0.0/ch12.pdf (Vowel sequences not to be used in Gurmukhi)

[5564] RFC 5564 (Code points to be excluded from repertoires for the Arabic language)

[6912] RFC 6912 (Code points considered problematic)

7. IANA Considerations

The IANA Services Operator is hereby requested to create the Registry of Unicode Code Points for Special Consideration in Network Identifiers, and to populate it with the values in section Section 5. The registry is to be updated by Expert Review.

This registry has no formal protocol status with respect to IDNA or PRECIS. It is a registry intended to be used by those creating registration or lookup policies, in order to inform the development of such policies.

8. Security Considerations

The registry established by this document is intended to help operators of identifier systems in deciding what to permit in identifiers. It may also be useful for user agents that attempt to provide warnings to users about suspicious or inadvisable identifiers. Operators that fail to make policies addressing the contents of the registry may permit the creation of identifiers that are misleading or that may be used in attacks on the network or users.

The registry is not a magic solution to all identifier ambiguity, and even refusing to permit registration of, or lookup of, every code point in the registry cannot ensure that misleading or confusing identifiers will never be created.

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4713] Lee, X., Mao, W., Chen, E., Hsu, N., and J. Klensin, "Registration and Administration Recommendations for Chinese Domain Names", RFC 4713, DOI 10.17487/RFC4713, October 2006, <https://www.rfc-editor.org/info/rfc4713>.

[RFC5890] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework", RFC 5890, DOI 10.17487/RFC5890, August 2010, <https://www.rfc-editor.org/info/rfc5890>.

[RFC5891] Klensin, J., "Internationalized Domain Names in Applications (IDNA): Protocol", RFC 5891, DOI 10.17487/RFC5891, August 2010, <https://www.rfc-editor.org/info/rfc5891>.

[RFC5892] Faltstrom, P., Ed., "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)", RFC 5892, DOI 10.17487/RFC5892, August 2010, <https://www.rfc-editor.org/info/rfc5892>.

[RFC5893] Alvestrand, H., Ed. and C. Karp, "Right-to-Left Scripts for Internationalized Domain Names for Applications (IDNA)", RFC 5893, DOI 10.17487/RFC5893, August 2010, <https://www.rfc-editor.org/info/rfc5893>.

[RFC5894] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Background, Explanation, and Rationale", RFC 5894, DOI 10.17487/RFC5894, August 2010, <https://www.rfc-editor.org/info/rfc5894>.

[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, <https://www.rfc-editor.org/info/rfc7564>.

[RFC7940] Davies, K. and A. Freytag, "Representing Label Generation Rulesets Using XML", RFC 7940, DOI 10.17487/RFC7940, August 2016, <https://www.rfc-editor.org/info/rfc7940>.

[UAX44] The Unicode Consortium, "Unicode Standard Annex #44, Unicode Character Database", <http://www.unicode.org/reports/tr44/>.

This references the most currently published version of the description of the Unicode Character Database.

[UCD] The Unicode Consortium, "Unicode Character Database", <http://www.unicode.org/Public/UCD/latest/ucd/>.

This references the most currently published version of the data files for the Unicode Character Database

[Unicode] The Unicode Consortium, "The Unicode Standard, Latest Version", <http://www.unicode.org/versions/latest/>.

This references the most currently published version

9.2. Informative References

[I-D.klensin-idna-5892upd-unicode70]
Klensin, J. and P. Faltstrom, "IDNA Update for Unicode 7.0 and Later Versions", draft-klensin-idna-5892upd-unicode70-05 (work in progress), October 2017.

[I-D.rfc5891bis]
Klensin, J., "Internationalized Domain Names in Applications (IDNA): Registry Restrictions and Recommendations", March 2017, <https://datatracker.ietf.org/doc/draft-klensin-idna-rfc5891bis/>.

[MSR] Integration Panel, ""Maximal Starting Repertoire (MSR-3)"", March 2018, <https://www.icann.org/en/system/files/files/msr-3-overview-28mar18-en.pdf>.

[RFC5564] El-Sherbiny, A., Farah, M., Oueichek, I., and A. Al-Zoman, "Linguistic Guidelines for the Use of the Arabic Language in Internet Domains", RFC 5564, DOI 10.17487/RFC5564, February 2010, <https://www.rfc-editor.org/info/rfc5564>.

[RFC6365] Hoffman, P. and J. Klensin, "Terminology Used in Internationalization in the IETF", BCP 166, RFC 6365, DOI 10.17487/RFC6365, September 2011, <https://www.rfc-editor.org/info/rfc6365>.

[RFC8228] Freytag, A., "Guidance on Designing Label Generation Rulesets (LGRs) Supporting Variant Labels", RFC 8228, DOI 10.17487/RFC8228, August 2017, <https://www.rfc-editor.org/info/rfc8228>.

[RZ-LGR] Integration Panel, ""Root Zone Label Generation Rules (LGR-2) - Overview and Summary"", July 2017, <https://www.icann.org/sites/default/files/lgr/lgr-2-overview-26jul17-en.pdf>.

Appendix A. Additional Background

A.1. The Theory of Inclusion

The mechanism that the IETF has come to prefer for internationalization of identifiers may be called "inclusion-based identifier internationalization", or "inclusion" for short. Under inclusion, the characters that are permissible in identifiers for a protocol are selected from the set of all Unicode characters. One starts with an empty set of characters, and then gradually adds characters to the set, usually based on Unicode properties (see below, and also Section 3).

Inclusion depends in part on assumptions the IETF made when the strategy was adopted and developed; some of those assumptions were about the relationships between different characters and the

likelihood that similar such relationships would get added to future versions of Unicode. Those assumptions turn out not to have been true in every case. Code points at issue are among those to be listed in the registry defined here. (See Section 5.)

The intent of Unicode is to encode all known writing systems into a single coded character set. One consequence of that goal is that Unicode encodes an enormous number of characters. Another is that the work of Unicode does not end until every writing system is encoded; even after that, it needs to continue to track any changes in those writing systems.

Unicode encodes abstract characters, not glyphs. Because of the way Unicode was built up over time, there are sometimes multiple ways to encode the same abstract character. For example, an e with an acute accent may be written by combining U+0065 LATIN SMALL LETTER E and U+0031 COMBINING ACUTE ACCENT, or it may be written U+00E9 LATIN SMALL LETTER E WITH ACUTE. If Unicode encodes an abstract character in more than one way, then for most purposes the different encodings should all be treated as though they’re the same character. This "canonical equivalence" between encodings of the same abstract characters is explicitly called out by Unicode. A lack of a defined canonical equivalence is tantamount to an assertion by Unicode that the two encodings do not represent the same abstract character, even if both happen to result in the same appearance.

Every encoded character in Unicode (more precisely, every code point) is associated with a set of properties. The properties define what script a code point is in, whether it is a letter or a number or punctuation and so forth, its direction when written, to what other code point or code point sequence it is canonically equivalent, and many other properties. These properties are important to the inclusion mechanism. They are defined in the Unicode Character Database [UCD] [UAX44].

Inclusion depends on the assumption that such strings as will be used in identifiers will not have any ambiguous matching to other strings. In practice, this means that input strings to the protocol are expected to be in Normalization Form C. This way, any alternative sequences of code points for the same characters will be normalized to a single form. If all the characters in the string are also included for the protocol’s candidate identifiers, then the string is eligible to be an identifier under the protocol.

A.2. The Difference Between Theory and Practice

In principle, under inclusion identifiers should be unambiguous. It has always been recognized, however, that for humans some ambiguity is inevitable, because of the vagaries of writing systems and of human perception.

Normalization Form C ("NFC") removes the ambiguities based on dual or multiple encoding for the same abstract character. However, characters are not the same as their glyphs. This means that it is possible for certain abstract characters to share a glyph. We can call such abstract characters "homoglyphs". While this looks at first like something that should be handled (or should have been handled) by normalization (NFC or something else), there are important differences; the situation is in some sense an extreme case of a spectrum of ambiguity.

A.2.1. Confusability

While Unicode deals in abstract characters and inclusion works on Unicode code points, users interact with strings as actually rendered: sequences of glyphs. There are characters that, depending on font, sometimes look quite similar to one another (such as "l" and "1"); any character that is like this is often called "visually similar". More difficult are characters that, in any normal rendering, always look the same as one another. The shared history of Cyrillic, Greek, and Latin scripts, for example, means that there are characters in each script that function similarly and that are usually indistinguishable from one another, though they are not the same abstract character. These are examples of "homoglyphs." Any character that can be confused for another one can be called confusable, and confusability can be thought of as a spectrum with "visually similar" at one end, and "homoglyphs" at the other. (We use the term "homoglyph" strictly: code points that normally use the same glyph when rendered.)

Note that homoglyphs are not restricted to cross-script scenarios - there are a number of homoglyphs where both code points or sequences are part of the same script.

A further issue is introduced by the fact that Unicode caters not only to living and dead languages alike, but also to scholarly and scientific notation, as well as specialized modes of written text, such as for poetry, religious works, or texts to be sung or chanted. Where these notations use symbols, they are excluded under inclusion, but where they use varieties of letter forms or marks used with letters, they are included by default. Some of these letters or marks, have been incorporated over time into orthographies for living

languages, which is one reason they were not rigorously excluded from the start. However, in some cases, they may (alone or in combination with ordinary letters appear the same (or very similar to) existing letters. This makes some of these characters, and especially the marks in question "troublesome".

Finally, IDNA 2008 has a limited appreciation for the fact that characters in complex scripts, unlike ASCII letters, cannot simply occur in random sequences. Neither software (for display or data entering) nor readers are prepared to process some of these code points "out of order". For such scripts, without a policy that describes permissible contexts, labels could be registered that cannot be rendered or typed reliably and which most users would not know how to read or recognize. In some cases, combining sequences typed in the "wrong" order may display identically to to those typed in the "correct" ordering; again something that needs to be sorted out by defining permissible contexts, for example by using the context rule mechanism in [RFC7940].

Appendix B. Examples

There are a number of cases that illustrate the combining sequence or digraph issue:

U+08A1 vs \u'0628'\u'0654' This case is ARABIC LETTER BEH WITH HAMZA ABOVE, which is the one that was detected during expert review that caused the IETF to first notice the issue, even though the issue existed before this. For detailed discussion of this case and some of the following ones, see [I-D.klensin-idna-5892upd-unicode70].

U+0681 vs \u'062D'\u'0654' This case is ARABIC LETTER HAH WITH HAMZA ABOVE, which (like U+08A1) does not have a canonical equivalent. In both cases, the places where hamza above and similar Arabic combining marks are used are specialized enough that the combining marks are generally excluded. See [RFC5564] and [RZ-LGR]. Unicode has a policy of encoding as composite any letter needed in an Arabic orthography, even if it appears superficially that the same shape could be achieved by a combining sequence. (In actual typography there's often a small but noticeable difference in placement of the mark between a composite character and a combining sequence.)

U+0623 vs \u'0627'\u'0654' This case is ARABIC LETTER ALEF WITH HAMZA ABOVE. Unlike the previous two cases, it does have a canonical equivalence with the combining sequence. Therefore, only the composite is used in IDNs.

U+09E1 vs u\’098C’u\’09E2’ This case is BENGALI LETTER VOCALIC LL. This is an example in the Bengali script of a case without a canonical equivalence to the combining sequence. Per Unicode, the single code point should be used to represent vowel signs in text, and the sequence of code points should not be used. There are similar cases in many Indic scripts. It is not a simple matter of disallowing the combining vowel mark in cases like this, because it is commonly used as vowel sign. The recommendation would be to add a context rule, restricting the vowel signs from appearing directly after an independent vowel like U+098C..

U+019A vs \u’006C’\u’0335’ This case is LATIN SMALL LETTER L WITH BAR. In at least some fonts, there is a detectable difference between the composite code point and the combining sequence, but only if one compares them side-by-side. Unlike a separable diacritic, there are no fast rules for placement of overlays. A bar may cross at different heights for different glyph shape or may cross different parts of the glyph. For this reason, there is no canonical equivalence defined between the sequence and the composite. Unicode has a principle of encoding barred letters of specific shape as single code point composites when needed for any writing system. The code point U+0335 COMBINING SHORT STROKE OVERLAY and similar overlay diacritics are therefore never needed as part of any orthography and are recommended to be excluded from identifiers.

U+00F8 vs \u’006F’\u’0337’ This is LATIN SMALL LETTER O WITH STROKE. The effect is similar to the previous case. Unicode has a principle of encoding stroked letters as composites when needed for any writing system.

U+02A6 vs \u’0074’\u’0073’ This is LATIN SMALL LETTER TS DIGRAPH, which is not canonically equivalent to the letters t and s. The intent appears to be that the digraph shows the two shapes as kerned, but the difference may be slight if viewed out of context. The use of the digraph is for specialized purposes; it can be excluded from identifiers.

U+01C9 vs \u’006C’\u’006A’ Unlike the TS digraph, the LJ digraph has a relevant compatibility decomposition, so it fails the relevant stability rules under inclusion and is therefore DISALLOWED in IDNA2008. This illustrates the way that consistencies that might be natural to some users of a script are not necessarily found in it, possibly because of uses by another writing system.

U+06C8 vs u\’0648’u\’0670’ ARABIC LETTER YU is an example where the normally-rendered character looks just like a combining sequence, but are named differently. This an example that shows that the

Unicode name is not a reliable indicator of the intended appearance. Like other cases in Arabig, the recommendation is to exclude the combining mark (and therefore the sequence) in favor of the composite.

U+069 vs \u'0069'\u'0307' LATIN SMALL LETTER I followed by COMBINING DOT ABOVE by definition, renders exactly the same as LATIN SMALL LETTER I by itself and does so in practice for any good font. The same would be true if "i" was replaced with any of the other Soft_Dotted characters defined in Unicode. The character sequence \u'0069'\u'0307' (followed by no other combining mark) is reportedly rather common on the Internet. Because base character and stand-alone code point are the same in this case, and the code points affected have the Soft_Dotted property already, this could be mitigated separately via a context rule affecting U+0307.

Other cases that demonstrate that the issue does not lie exclusively or primarily with combining sequences:

U+0B95 vs U+0BE7 The TAMIL LETTER KA and TAMIL DIGIT ONE are always indistinguishable, but needed to be encoded separately because one is a letter and the other is a digit.

Arabic-Indic Digits vs. Extended Arabic-Indic Digits Seven digits of these two sequences have entirely identical shapes. This case is an example of something dealt with in inclusion that nevertheless can lead to confusions that are not fully mitigated. IDNA, for example, contains context rules restricting the digits to one set or another; but such rules apply only to a single label, not to an entire name. Moreover, it provides no way of distinguishing between two labels that both conform to the context rule, but where each contains a different member one of the seven identical shape pairs.

U+53E3 vs U+56D7 These are two Han characters (roughly rectangular) that are different when laid side by side; but they may be difficult to distinguish out of context or in very small print.

U+01DD vs U+0259 The two Latin script code points share the have the identical appearance of a lower-case upside down "e". They are encoded differently due to different uppercase forms. The fact that they uppercase differently is taken as evidence that they are not the same abstract character, despite the superficial evidence of their shared shape. The more common cases, where the uppercase forms are identical may be of less concern, given that IDNA 2008 is limited to lower case.

Cross script homoglyphs usually do not involve combining sequences, but can be mitigated by rules requiring strings to be in a single script. For zones that support multiple scripts, it may be necessary to have policies to prevent whole-script homographs: labels entirely in one script that look the same as another label in the other script. One method would be to define "blocked" variants (See [RFC7940] and [RFC8228]).

LATIN SMALL LETTER OPEN E is one of a handful of examples of characters borrowed from another script, in this case GREEK SMALL LETTER EPSILON.

LATIN SMALL LETTER E and CYRILLIC SMALL LETTER IE are historically related, both derive from uppercase forms of the GREEK CAPTIAL LETTER EPSILON. There are a number of such pairs -- enough to make many whole strings that look the same in both scripts (but usually spell nonsense in one of them). An example would be "pax".

Appendix C. Discussion Venue

Note to RFC Editor: this section should be removed prior to publication as an RFC.

This Internet-Draft may be discussed on the IAB Internationalization public list: i18n-discuss@iab.org.

Appendix D. Change History

Note to RFC Editor: this section should be removed prior to publication as an RFC.

00:

* Initial version

01:

* Add background and examples from the LUCID Problem Statement
* Add a paragraph about motivation to explain the difference between this registry and administrative policy more generally
* Expand and clarify a number of earlier points of discussion
* Attempt to make clear that this registry does not update any protocols

* Move some formerly-appendix material to the body

* Expand the initial registry.

02:

* Expanded the discussion of possible mitigation approaches and made its own section.

* Added more detail to the categories of troublesome characters

* Minor updates to "Existing techniques" section.

* Some extension to the description of the contents of the registry and discussion of how to handle additional information.

Authors' Addresses

Asmus Freytag
ASMUS, Inc.

Email: asmus@unicode.org

John C Klensin
1770 Massachusetts Ave, Ste 322
Cambridge, MA 02140
U.S.A.

Email: john-ietf@jck.com

Andrew Sullivan
Oracle Corp.
100 Milverton Drive
Missisauga, ON L5R 4H1
Canada

Email: andrew.s.sullivan@oracle.com

Network Working Group
Internet-Draft
Updates: 5892, 5894 (if approved)
Intended status: Standards Track
Expires: April 11, 2018

J. Klensin

P. Faltstrom
Netnod

October 8, 2017

IDNA Update for Unicode 7.0 and Later Versions
draft-klensin-idna-5892upd-unicode70-05

Abstract

The current version of the IDNA specifications anticipated that each new version of Unicode would be reviewed to verify that no changes had been introduced that required adjustments to the set of rules and, in particular, whether new exceptions or backward compatibility adjustments were needed. The review for Unicode 7.0.0 first identified a potentially problematic new code point and then a much more general and difficult issue with Unicode normalization. This specification discusses those issues and proposes updates to IDNA and, potentially, the way the IETF handles comparison of identifiers more generally, especially when there is no associated language or language identification. It also applies an editorial clarification to RFC 5892 that was the subject of an earlier erratum and updates RFC 5894 to point to the issues involved.

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 https://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 11, 2018.

Copyright Notice

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction

Note in/about -04 and -05 Drafts: These two versions of the document contains a very large amount of new material as compared to the -03 version. The new material reflects an evolution of community understanding in the first quarter of 2015 and further evolution between then and mid-2017 from an assumption that the problem involved only a few code points and one combining character in a single script (Hamza Above and Arabic) to an understanding that the problem we have come to call "non-decomposing code points" and several closely related ones are quite pervasive and may represent fundamental misunderstandings or omissions from IDNA2008 (and, by extension, the basics of PRECIS [RFC8264]) that must be corrected if those protocols are going to be used in a way that supports internationalized identifiers on the Internet predictably (as seen by the end user) and securely.

This version is still necessarily incomplete: not only is our understanding probably still not comprehensive, but there are a number of placeholders for text and references. Nonetheless, the document in its current form should be useful as both the beginning of a comprehensive overview is the issues and a source of references to other relevant materials.

This draft could almost certainly be better organized to improve its readability: specific suggestions would be welcome.

1.1. Origins and Discovery of the Issue

The current version of the IDNA specifications, known as "IDNA2008" [RFC5890], anticipated that each new version of Unicode would be reviewed to verify that no changes had been introduced that required adjustments to IDNA's rules and, in particular, whether new exceptions or backward compatibility adjustments were needed. When that review was carefully conducted for Unicode 7.0.0 [Unicode7], comparing it to prior versions including the text in Unicode 6.2 [Unicode62], it identified a problematic new code point (U+08A1, ARABIC LETTER BEH WITH HAMZA ABOVE). The code point was added for Arabic Script use with the Fula (also known as Fulfulde, Pulaar, amd Pular'Fulaare) language. That language is apparently most often written in Latin characters today [Omniglot-Fula] [Dalby] [Daniels].

The specific problem is discussed in detail in Section 3. In very broad terms, IDNA (and other IETF work) assume that, if one can represent "the same character" either as a combining sequence or as a single code point, strings that are identical except for those alternate forms will compare equal after normalization. Part of the difficulty that has characterized this discussion is that "the same" differs depending on the criteria that are chosen. It may be further complicated in practice by differences in preferred type styles or rendering, but Unicode code point choices are not supposed to depend on type style (font) variations and, again, IDNA has no mechanism for specifying language choices that might affect rendering.

The behavior of the newly-added code point, while non-optimal for IDNA, follows that of a few code points that predate Unicode 7.x and even the IDNA 2008 specifications and Unicode 6.0. Those existing code points, which may not be easy to accurately characterize as a group, make the question of what, if anything, to do about this new exceedingly problematic one and, perhaps separately, what to do about existing sets of code points with the same behavior, because different reasonable criteria yield different decisions, specifically:

- To disallow it (and future, but not existing, characters with similar characteristics) as an IDNA exception case creates inconsistencies with how those earlier code points were handled.

- To disallow it and the similar code points as well would necessitate invalidating some potential labels that would have been valid under IDNA2008 until this time. Depending on how the

collection of similar code points is characterized, a few of them are almost certainly used in reasonable labels.

- To permit the new code point to be treated as PVALID creates a situation in which it is possible, within the same script, to compose the same character symbol (glyph or grapheme) in two different ways that do not compare equal even after normalization. That condition would then apply to it and the earlier code points with the same behavior. That situation contradicts a fundamental assumption of IDNA that is discussed in more detail below.

NOTE IN DRAFT:

This working draft discusses six alternatives, including an idea (an IETF-specific normalization form) that seemed too drastic to be considered when IDNA2008 was designed or even when the review of Unicode 7.0 for IDAN purposes began. In retrospect, it not only would have been appropriate to discuss when the IDNA2008 specifications were being developed but is appearing more attractive now. The authors suggest that the community discuss the relevant tradeoffs and make a decision and that the document then be revised to reflect that decision, with the other alternatives discussed as options not chosen. Because there is no ideal choice, the discussion of the issues in Section 3 is probably as or more important than the particular choice of how to handle this code point. In addition to providing information for this document, that section should be considered as an updating addendum to RFC 5894 [RFC5894] and should be incorporated into any future revision of that document.

As the result of this version of the document containing several alternate proposals, some of the text is also a little bit redundant. That will be corrected in future versions.

1.2. IDNA2008 and Special or Exceptional Cases

IDNA2008 contains several type of explicit provisions for characters (code points) that require special treatment when the requirements of the DNS cannot easily be met by calculations based on stable Unicode properties. Those provisions are
[[CREF1: ... to be supplied]]

As anticipated when IDNA2008, and RFC 5892 in particular, were written, exceptions and explicit updates are likely to be needed only if there is disagreement between the Unicode Consortium’s view about what is best for the Standard and its very diverse user community and the IETF’s view of what is best for IDNs, the DNS, and IDNA. It was hoped that a situation would never arise in which the the two

perspectives would disagree, but the possibility was anticipated and considerable mechanism added to RFC 5890 and 5982 as a result. It is probably important to note that a disagreement in this context does not imply that anyone is "wrong", only that the two different groups have different needs and therefore criteria about what is acceptable. In particular, it appears that the Unicode Consortium has made assumptions about the availability (by explicit designation or context) of information about applicable languages or other context for a give string that are not possible for IDNA. For that reason, the IETF has, in the past, allowed some characters for IDNA that active Unicode Technical Committee members suggested be disallowed to avoid a change in derived tables [RFC6452]. This document describes a set of cases for which the IETF must consider disallowing sets of characters that the various properties would otherwise treat as PVALID.

This document provides the "flagging for the IESG" specified by Section 5.1 of RFC 5892. As specified there, the change itself requires IETF review because it alters the rules of Section 2 of that document.

[[RFC Editor: please remove the following comment and note if they get to you.]]

[[IESG: It might not be a bad idea to incorporate some version of the following into the Last Call announcement.]]

NOTE IN DRAFT to IETF Reviewers: The issues in this document, and particularly the choices among options for either adding exception cases to RFC 5892 or ignoring the issue, warning people, and hoping the results do not include or enable serious problems, are fairly esoteric. Understanding them requires that one have at least some understanding of how scripts in which precomposed characters are preferred over combining sequences as a Unicode design and extension principle work. Those scripts include Arabic but, unlike the assumption when the issues were first discovered, are by no means limited to it. Readers should also understand the reasons the Unicode Standard gives various Arabic Script characters a fairly extended discussion [Unicode70-Arabic] but should treat that only as an example and note that most other cases are much less well documented. It also requires understanding of a number of Unicode principles, including the Normalization Stability rules [UAX15-Versioning] as applied to new precomposed characters and guidelines for adding new characters. There is considerable discussion of the issues in Section 3 and references are provided for those who want to pursue them, but potential reviewers should assume that the background needed to understand the reasons for this change is no less deep in the

subject matter than would be expected of someone reviewing a proposed change in, e.g., the fundamentals of BGP, TCP congestion control, or some cryptographic algorithm. Put more bluntly, one's ability to read or speak languages other than English, or even one or more languages that use the Arabic script or other scripts similarly affected, does not make one an expert in these matters.

1.3. Terminology

This document assumes that the reader is reasonably familiar with the terminology of IDNA [RFC5890] and Unicode [Unicode7] and with the IETF conventions for representing Unicode code points [RFC5137]. Some terms used here may not be used in the same way in those two sets of documents. From one point of view, those differences may have been the results of, or led to, misunderstandings that may, in turn, be part of the root cause of the problems explored in this document. In particular, this document uses the term "precomposed character" to describe characters that could reasonably be composed by a combining sequence using code points with appropriate appearance in common type styles but for which a single code point that does not require combining sequences is available. That definition is strictly about mechanical composition and does not involve any considerations about how the character is used. It is closely related to this document's definition of "identical". When a precomposed character exists and either applying NFC to the combining sequence does not yield that character or applying NFD to that character's code point does not yield the combining sequence, it is referred to in this document as "non-decomposable".

The document also uses some terms that are familiar to those who have been involved with IDNs and IDNA for a long time, but uses them more precisely than may be common in other quarters. For example, the term "Punycode" is not used at all in the rest of this document because it is the name of a very specific encoding algorithm [RFC3492] that does not incorporate the rules and algorithms for domain name labels that are produced by that encoding. Instead, the generic terms "ACE" or "ACE string" for "ASCII-compatible encoding" is used to refer to strings that abstractly contain characters outside the ASCII repertoire [RFC0020] but are encoded so that only ASCII characters appear in the string that would be encountered by a user or protocol and the terms "A-label" and "U-label", as defined in RFC 5890, to refer to the ACE and more conventional (or "native") character forms in which those non-ASCII characters appear in conventional Unicode encodings (typically UTF-8).

2. Document Aspirations

This document, in its present form, is not a proposal for a solution. Instead, it is intended to be (or evolve into) a comprehensive description of the issues and problems and to outline some possible approaches to a solution. A perfect solution -- one that would resolve all of the issues identified in this document -- would involve a relatively small set of relatively simple rules and hence would be comprehensible and predictable for and by non-expert end users, would not require code point by code point or even block by block exception lists, and would not leave uses of any script or language feeling that their particular writing system have been treated less fairly than others.

Part of the reality we need to accept is that IDNA, in its present form, represents compromises that does not completely satisfy those criteria and whatever is done about these issues will probably make it (or the job of administering zones containing IDNs) more complex. Similarly, as the Unicode Standard suggests when it identifies ten Design Principles and the text then says "Not all of these principles can be satisfied simultaneously..." [Unicode70-Design], while there are guidelines and principles, a certain amount of subjective judgment is involved in making determinations about normalization, decomposition, and some property values. For Unicode itself, those issues are resolved by multiple statements (at least one cited below) that one needs to rely on per-code point information in the Unicode Character Database rather than on rules or principles. The design of IDNA and the effort to keep it largely independent of Unicode versions requires rules, categories, and principles that can be relied upon and applied algorithmically. There is obviously some tension between the two approaches.

3. Problem Description

3.1. IDNA assumptions about Unicode normalization

IDNA makes several assumptions about Unicode, Unicode "characters", and the effects of normalization. Those assumptions were based on careful reading of the Unicode Standard at the time [Unicode5], guided by advice and commitments by members of the Unicode Technical Committee. Those assumptions, and the associated requirements, are necessitated by three properties of DNS labels that typically do not apply to blocks of running text:

1. There is no language context for a label. While particular DNS zones may impose restrictions, including language or script restrictions, on what labels can be registered, neither the DNS nor IDNA impose either type of restriction or give the user of a

label any indication about the registration or other restrictions that may have been imposed.

2. Labels are often mnemonics rather than words in any language. They may be abbreviations or acronyms or contain embedded digits and have other characteristics that are not typical of words.

3. Labels are, in practice, usually short. Even when they are the maximum length allowed by the DNS and IDNA, they are typically too short to provide significant context. Statements that suggest that languages can almost always be determined from relatively short paragraphs or equivalent bodies of text do not apply to DNS labels because of their typical short length and because, as noted above, they are not required to be formed according to language-based rules.

At the same time, because the DNS is an exact-match system, there must be no ambiguity about whether two labels are equal. Although there have been extensive discussions about "confusingly similar" characters, labels, and strings, such tests between scripts are always somewhat subjective: they are affected by choices of type styles and by what the user expects to see. In spite of the fact that the glyphs that represent many characters in different scripts are identical in appearance (e.g., basic Latin "a" (U+0061) and the identical-appearing Cyrillic character (U+0430), the most important test is that, if two glyphs are the same within a given script, they must represent the same character no matter how they are formed.

Unicode normalization, as explained in [UAX15], is expected to resolve those "same script, same glyph, different formation methods" issues. Within the Latin script, the code point sequence for lower case "o" (U+006F) and combining diaeresis (U+0308) will, when normalized using the "NFC" method required by IDNA, produce the precomposed small letter o with diaeresis (U+00F6) and hence the two ways of forming the character will compare equal (and the combining sequence is effectively prohibited from U-labels).

NFC was preferred over other normalization methods for IDNA because it is more compact, more likely to be produced on keyboards on which the relevant characters actually appeared, and because it does not lose substantive information (e.g., some types of compatibility equivalence involves judgment calls as to whether two characters are actually the same -- they may be "the same" in some contexts but not others -- while canonical equivalence is about different ways to produce the glyph for the same abstract character).

IDNA also assumed that the extensive Unicode stability rules would be applied and work as specified when new code points were added. Those

rules, as described in The Unicode Standard and the normative annexes identified below, provide that:

1. New code points representing precomposed characters that can be formed from combining sequences will not be added to Unicode unless neither the relevant base character nor required combining character(s) are part of the Standard within the relevant script [UAX15-Versioning].

2. If circumstances require that principle be violated, normalization stability requires that the newly-added character decompose (even under NFC) to the previously-available combining sequence [UAX15-Exclusion].

At least at the time IDNA2008 was being developed, there was no explicit provision in the Standard's discussion of conditions for adding new code points, nor of normalization stability, for an exception based on different languages using the same script or ambiguities about the shape or positioning of combining characters.

3.2. The discovery and the Arabic script cases

While the set of problems with normalization discussed above were discovered with a newly-added code point for the Arabic Script and some characteristics of Unicode handling of that script seem to make the problem more complex going forward, these are not issues specific to Arabic. This section describes the Arabic-specific problems; subsequent ones (starting with Section 3.3) discuss the problem more generally and include illustrations from other scripts.

3.2.1. New code point U+08A1, decomposition, and language dependency

Unicode 7.0.0 introduces the new code point U+08A1, ARABIC LETTER BEH WITH HAMZA ABOVE. As can be deduced from the name, it is visually identical to the glyph that can be formed from a combining sequence consisting of the code point for ARABIC LETTER BEH (U+0628) and the code point for Combining Hamza Above (U+0654). The two rules summarized above (see the last part of Section 3.1) suggest that either the new code point should not be allocated at all or that it should have a decomposition to \u'0628'\u'0654'.

Had the issues outlined in this document been better understood at the time, it probably would have been wise for RFC 5892 to disallow either the precomposed character or the combining sequence of each pair in those cases in which Unicode normalization rules do not cause the right thing to happen, i.e., the combining sequence and precomposed character to be treated as equivalent. Failure to do so at the time places an extra burden on registries to be sure that

conflicts (and the potential for confusion and attacks) do not exist. Oddly, had the exclusion been made part of the specification at that time, the preference for precomposed forms noted above would probably have dictated excluding the combining sequence, something not otherwise done in IDNA2008 because the NFC requirement serves the same purpose. Today, the only thing that can be excluded without the potential disruption of disallowing a previously-PVALID combining sequence is the to exclude the newly-added code point so whatever is done, or might have been contemplated with hindsight, will be somewhat inconsistent.

3.2.2. Other examples of the same behavior within the Arabic Script

One of the things that complicates the issue with the new U+08A1 code point is that there are several other Arabic-script code points that behave in the same way for similar language-specific reasons.

In particular, at least three other grapheme clusters that have been present for many version of Unicode can be seen as involving issues similar to those for the newly-added ARABIC LETTER BEH WITH HAMZA ABOVE. ARABIC LETTER HAH WITH HAMZA ABOVE (U+0681) and ARABIC LETTER REH WITH HAMZA ABOVE (U+076C) do not have decomposition forms and are preferred over combining sequences using HAMZA ABOVE (U+0654) [Unicode70-Hamza]. By contrast, ARABIC LETTER ALEF WITH HAMZA ABOVE (U+0623) decomposes into \u’0627’\u’0654’, ARABIC LETTER WAW WITH HAMZA ABOVE (U+0624) decomposes into \u’0648’\u’0654’, and ARABIC LETTER YEH WITH HAMZA ABOVE (U+0626) decomposes into \u’064A’\u’0654’ so the precomposed character and combining sequences compare equal when both are normalized, as this specification prefers.

There are other variations in which a precomposed character involving HAMZA ABOVE has a decomposition to a combining sequence that can form it. For example, ARABIC LETTER U WITH HAMZA ABOVE (U+0677) has a compatibility decomposition. but not a canonical one, into the combining sequence \u’06C7’\u’0674’.

3.2.3. Hamza and Combining Sequences

As the Unicode Standard points out at some length [Unicode70-Arabic], Hamza is a problematic abstract character and the "Hamza Above" construction even more so [Unicode70-Hamza]. Those sections explain a distinction made by Unicode between the use of a Hamza mark to denote a glottal stop and one used as a diacritic mark to denote a separate letter. In the first case, the combining sequence is used. In the second, a precomposed character is assigned.

Unlike Unicode generally and because of concerns about identifier spoofing and attacks based on similarities, character distinctions in

IDNA are based much more strictly on the appearance of characters; language and pronunciation distinctions within a script are not considered. So, for IDNA, BEH WITH HAMZA ABOVE is not-quite-tautologically the same as BEH WITH HAMZA ABOVE, even if one of them is written as U+08A1 (new to Unicode 7.0.0) and the other as the sequence \u'0628'\u'0654' (feasible with Unicode 7.0.0 but also available in versions of Unicode going back at least to the version [Unicode32] used in the original version of IDNA [RFC3490]. Because the precomposed form and combining sequence are, for IDNA purposes, the same, IDNA expects that normalization (specifically the requirement that all U-labels be in NFC form) will cause them to compare equal.

If Unicode also considered them the same, then the principle would apply that new precomposed ("composition") forms are not added unless one of the code points that could be used to construct it did not exist in an earlier version (and even then is discouraged) [UAX15-Versioning]. When exceptions are made, they are expected to conform to the rules and classes in the "Composition Exclusion Table", with class 2 being relevant to this case [UAX15-Exclusion]. That rule essentially requires that the normalization for the old combining sequence to itself be retained (for stability) but that the newly-added character be treated as canonically decomposable and decompose back to the older sequence even under NFC. That was not done for this particular case, presumably because of the distinction about pronunciation modifiers versus separate letters noted above. Because, for IDNA and the DNS, there is a possibility that the composing sequence \u'0628'\u'0654' already appears in labels, the only choice other than allowing an otherwise-identical, and identically-appearing, label with U+08A1 substituted to identify a different DNS entry is to DISALLOW the new character.

3.3. Precomposed characters without decompositions more generally

3.3.1. Description of the general problem

As mentioned above, IDNA made a strong assumption that, if there were two ways to form the same abstract character in the same script, normalization would result in them comparing equal. Work on IDNA2008 recognized that early version of Unicode might also contain some inconsistencies; see Section 3.3.2.3.2 below.

Having precomposed code points exist that don't have decompositions, or having code points of that nature allocated in the future, is problematic for those IDNA assumptions about character comparison. It seems to call for either excluding some set of code points that IDNA's rules do not now identify, development and use of a normalization procedure that behaves as expected (those two options

may be nearly equivalent for many purposes), or deciding to accept a risk that, apparently, will only increase over time.

It is not clear whether the reasons the IDNABIS WG did not understand and allow for these cases are important except insofar as they inform considerations about what to do in the future. It seemed (and still seems to some people) that the Unicode Standard is very clear on the matter (or at least was when IDNA2008 was being developed). In addition to the normalization stability rules cited in the last part of Section 3.1. the discussion in the Core Standard seems quite clear. For example, "Where characters are used in different ways in different languages, the relevant properties are normally defined outside the Unicode Standard" in Section 2.2, subsection titled "Semantics" [Unicode7] did not suggest to most readers that sometimes separate code points would be allocated within a script based on language considerations. Similarly, the same section of the Standard says, in a subsection titled "Unification", "The Unicode Standard avoids duplicate encoding of characters by unifying them within scripts across language" and does not list exceptions to that rule or limit it to a single script although it goes on to list "CJK" as an example. Another subsection, "Equivalent Sequences" indicates "Common precomposed forms ... are included for compatibility with current standards. For static precomposed forms, the standard provides a mapping to an equivalent dynamically composed sequence of characters". The latter appears to be precisely the "all precomposed characters decompose into the relevant combining sequences if the relevant base and combining characters exist in the Standard" rule that IDNA needs and assumed and, again, there is no mention of exceptions, language-dependent of otherwise. The summary of stability policies cited in the Standard [Unicode70-Stability] does not appear to shed any additional light on these issues.

The Standard now contains a subsection titled "Non-decomposition of Overlaid Diacritics" [Unicode70-Overlay] that identifies a list of diacritics that do not normally form characters that have decompositions. The rule given has its own exceptions and the text clearly states that there is actually no way to know whether a code point has a decomposition other than consulting the Unicode Character Database entry for that code point. The subsequent section notes that this can be a security problem. While the issues with IDNA go well beyond what is normally considered security, that comment now seems clear. While that subsection is helpful in explaining the problem, especially for European scripts, it does not appear in the Unicode versions that were current when IDNA2008 was being developed.

3.3.2. Latin Examples and Cases

While this set of problems was discovered because of a code point added to the Arabic script in precombined form to support a particular language, there are actually far more examples for, e.g., Latin script than there are for Arabic script. Many of them are associated with the "non-decomposition of combining diacriticals" issues mentioned above, but the next subsections describe other cases that are not directly bound to decomposition.

3.3.2.1. The font exclusion and compatability relationships

Unicode contains a large collection of characters that are identified as "Mathematical Symbols". A large subset of them are basic or decorated Latin characters, differing from the ordinary ones only by their usage and, in appearance, by font or type styling (despite the general principle that font distinctions are not used as the basis for assigning separate code points. Most of these have canonical mappings to the base form, which eliminates them from IDNA, but others do not and, because the same marks that are used as phonetic diacritical markings in conventional alphabetical use have special mathematical meanings, applications that permit the use of these characters have their own issues with normalization and equality.

3.3.2.2. The phonetic notation characters and extensions

Another example involves various Phonetic Alphabet and Extension characters. many of which, unlike the Mathematical ones, do not have normalizations that would make them compare equal to the basic characters with essentially identical representations. This would not be a problem for IDNA if they were identified with a specialized script or as symbols rather than letters, but neither is the case: they are generally identified as lower case Latin Script letters even when they are visually upper-case, another issue for IDNA.

3.3.2.3. The stroke (solidus) ambiguity

Some combining characters have two or more forms. for example, in the case of the character popularly known as "slash", "stroke", or "solidus" (sometime prefixed by "forward"), there are "short" and "long" combining forms, U+0337 (COMBINING SHORT SOLIDUS OVERLAY) and U+0338 (COMBINING LONG SOLIDUS OVERLAY). It is not clear how long a short one needs to be to make it "long" or how short a long one needs to be to make it "short". Perhaps for that reason, U+00F8 has no decomposition and neither U+006F U+0337 nor U+006F U+0338 combine to it with NFC.

Adding to the confusion, at least when one attempts to use Unicode character names to identify places to look for problems, U+00F8 is formally called LATIN SMALL LETTER O WITH STROKE but, in combining character terminology, the term "stroke" refers to a horizontal bar, not an angled one, as in U+0335 and U+0336 (also short and long versions). However, when one overlays one of those on an "o" (U+006F), one gets U+0275, LATIN SMALL LETTER BARRED O, not "...o with stroke". That character, by the way, does not decompose either. This does illustrate the principle that it is not feasible to rely on Unicode code point names to identify confusable character sequences, even ones that produce the same, more or less font-independent, grapheme clusters.

3.3.2.3.1. Combining dots and other shapes combine... unless...

The discussion of "Non-decomposition of Overlaid Diacritics" [Unicode70-Overlay] indirectly exhibits at least one reason why it has been difficult to characterize the problem. If one combines that subsection with others, one gets a set of rules that might be described as:

1. If the precomposed character and the code points that make up the combining sequence exist, then canonical composition and decomposition work as expected, except...

2. If the precomposed character was added to Unicode after the code points that make up the combining sequence, normalization stability for the combining sequences requires that NFC applied to the precomposed character decomposes rather than having the combining sequence compose to the new character, however...

3. If the combining sequence involves a diacritic or other mark that actually touches the base character when composed, the precomposed character does not have a decomposition, unless...

4. The combining diacritic involved is Cedilla (U+0327), Ogonek (U+0328), or Horn (U+031B), in which case the precomposed characters that contain them "regularly" (but presumably not always) decomposes, and...

5. There are further exceptions for Hamza which does not overlay the associated base character in the same way the Latin-derived combining diacritics and other marks do. Those decisions to decompose a precomposed character (or not) are based on language or phonetic considerations, not the combining mechanism or appearance, or perhaps,...

6. Some characters have compatibility decompositions rather than canonical ones [Unicode70-CompatDecomp]. Because compatibility relationships are treated differently by IDNA, PRECIS [RFC8264], and, potentially, other protocols involving identifiers for Internet use, the existence of compatibility relationship may or may not be helpful. Finally,...

7. There is no reason to believe the above list is complete. In particular, if whether a precomposed character decomposes or not is determined by language or phonetic distinctions or by a decision that all new characters for some scripts will be precomposed while new ones for others will be added (if needed) as combining sequences, one may need additional rules on a per-script and/or per-character basis.

The above list only covers the cases involving combining sequences. It does not cover cases such as those in Section 3.3.2.1 and Section 3.3.2.2 and there may be additional groups of cases not yet identified.

3.3.2.3.2. "Legacy" characters and new additions

The development of categories and rules for IDNA recognized that early version of Unicode might contain some inconsistencies if evaluated using more contemporary rules about code point assignments and stability. In particular, there might be some exceptions from different practices in early version of Unicode or anomalies caused by copying existing single- or dual-script standards into Unicode as block rather than individual character additions to the repertoire. The possibility of such "legacy" exceptions was one reason why the IDNA category rules include explicit provisions for exception lists (even though no such code points were identified prior to 2014).

3.3.3. Unexpected Combining Sequances

Most combining characters have the script property "Inherited" or "Common", i.e., are not members of any particular script and will not cause rules against mixed-script labels to be triggered. Normalization rules are generally structured around the base character, so unexpected combinations of base characters with combining ones may lead to cases where normalization might normally be expected to produce a precombined character but does not do so (in the most common situation because no such precombined character exists. For example, the Latin script characters "a" and "a with acute accent" are both coded (as U+0061 and U+00E1). If the latter is coded as the combining sequence U+0061 U+0301, NFC will turn that sequence into U+00E1 and everything will work as users expect. However, the Cyrillic "a" character (U+0430) is notoriously similar

in appearance in most type styles to U+0061 and the U+0439 U+0301 and that sequence does not normalize to anything else. Because thre is no code point assigned for Cyrillic small letter a with acute accent and unlike many of the other examples in this document, that is Unicode working exactly as would be expected. Whether it is an issue or not depends on the questions that are being asked and what rules are being applied.

3.3.4. Examples and Cases from Other Scripts

Research into these issues has not yet turned up a comprehensive list of affected scripts and code points. As discussed elsewhere in this document, it is clear that Arabic and Latin Scripts are significantly affected, that some Han and Kangxu radicals and ideographs are affected, and that other examples do exist -- it is just not known how many of those examples there are and what patterns, if any, characterize them.

3.3.4.1. Scripts with precomposed preferences and ones with combining preferences

While the authors have been unable to find an explanation for the differentiation in the Unicode Standard, we have been told that there are differences among scripts as to whether the action preference is to add new combining sequences only (and resist adding precomposed characters) as suggested in Section 3.3.2.3.1 or to add precomposed characters, often ones that do not have decompositions. If those difference in preference do exist, it is probably important to have them documented so that they can be reflected in IDNA review procedures and elsewhere. It will also require IETF discussion of whether combining sequences should be deprecated when the corresponding precomposed characters are added or to disallow combining sequences entirely for those scripts (as has been implicitly suggested for Arabic language use [RFC5564]).

[[CREF2: The above isn’t quite right and probably needs additional discussion and text.]]

3.3.4.2. The Han and Kangxu Cases

[[CREF3: .. to be supplied ..]]

3.4. Confusion and the Casual User

To the extent to which predictability for relatively casual users is a desired and important feather of relevant application or application support protocols, it is probably worth observing that the complex of rules and cases suggested or implied above is almost

certainly too involved for the typical such user to develop a good intuitive understanding of how things behave and what relationships exist. Conversely, the nature of writing systems for natural languages, especially those that have evolved and diverged over centuries, implies that no set of rules about allowable characters will guarantee complete safety (however that is defined).

4. Implementation options and issues: Unicode properties, exceptions, and the nature of stability

4.1. Unicode Stability compared to IETF (and ICANN) Stability

The various stability rules in Unicode [Unicode70-Stability] all appear to be based on the model that once a value is assigned, it can never be changed. That is probably appropriate for a character coding system with multiple uses and applications. It is probably the only option when normative relationships are expressed in tables of values rather than by rules. One consequence of such a model is that it is difficult or impossible to fix mistakes (for some stability rules, the Unicode Standard does provide for exceptions) and even harder to make adjustments that would normally be dictated by evolution.

"No changes" provides a very strong and predictable type of stability. There are many reasons to take that path. As in some of the cases that motivated this document, the difficulty is that simply adding new code points (in Unicode) or features (in a protocol or application) may be destabilizing. One then has complete stability for systems that never use or allow the new code points or features, but rough edges for newer systems that see the discrepancies and rough edges. IDNA2003 (inadvertently) took that approach by freezing on Unicode 3.2 -- if no code points added after Unicode 3.2 had ever been allowed, we would have had complete stability even as Unicode libraries changed. Unicode has been quite ingenious about working around those difficulties with such provisions as having code points for newly-added precomposed characters decompose rather than altering the normalization for the combining sequences. Other cases, such as newly-added precomposed characters that do not decompose for, e.g., language or phonetic reasons, are more problematic.

The IETF (and ICANN and standards development bodies such as ISO and ISO/IEC JTC1) have generally adopted a different type of stability model, one which considers experience in use and the ill effects of not making changes as well as the disruptive effects of doing so. In the IETF model, if an earlier decision is causing sufficient harm and there is consensus in the communities that are most affected that a change is desirable enough to make transition costs acceptable, then the change is made.

The difference and its implications are perhaps best illustrated by a disagreement when IDNA2008 was being approved. IDNA2003 had effectively prevented some characters, notably (measured by intensity of the protests) the Sharp S character (U+00DF) from being used in DNS labels by mapping them to other characters before conversion to ACE form. It has also prohibited some other code points, notably ZWJ (U+200D) and ZWNJ (U+200C), by discarding them. In both cases, there were strong voices from the relevant language communities, supported by the registry communities, that the characters were important enough that it was more desirable to undergo the short-term pain of a transition and some uncertainty than to continue to exclude those characters and the IDNA2008 rules and repertoire are consistent with that preference. The Unicode Consortium apparently believed that stability --elimination of any possibility of label invalidation or different interpretations of the same string-- was more important than those writing system requirements and community preferences. That view was expressed through what was effectively a fork in (or attempt to nullify) the IETF Standard [UTS46] a result that has probably been worse for the overall Internet than either of the possible decision choices.

4.2. New Unicode Properties

One suggestion about the way out of these problems would be to create one or more new Unicode properties, maintained along with the rest of Unicode, and then incorporated into new or modified rules or categories in IDNA. Given the analysis in this document, it appears that that property (or properties) would need to provide:

1. Identification of combining characters that, when used in combining sequences, do not produce decomposable characters. [[CREF4: Wording on the above is not quite right but, for the present, maybe the intent is clear.]]

2. Identification of precomposed characters that might reasonably be expected to decompose, but that do not.

3. Identification of character forms that are distinct only because of language or phonetic distinctions within a script.

4. Identification of scripts for which precomposed forms are strongly preferred and combining sequences should either be viewed as temporary mechanisms until precomposed characters are assigned or banned entirely.

5. Identification of code points that represent symbols for specific, non-language, purposes even if identified as letters or numerals by their General Property. This would include all

characters given separate code points because of specialized "mathematical" and "phonetic" characters (see Section 3.3.2.2 and Section 3.3.2.1), but there are probably additional cases.

Some of these properties (or characteristics or values of a single property) would be suitable for disallowing characters, code points, or contextual sequences that otherwise might be allowed by IDNA. Others would be more suitable for making equality comparisons come out as needed by IDNA, particularly to eliminate distinctions based on language context.

While it would appear that appropriate rules and categories could be developed for IDNA (and, presumably, for PRECIS, etc.) if the problem areas are those identified in this document, it is not yet known whether the list is complete (and, hence, whether additional properties or information would be needed).

Even with such properties, IDNA would still almost certainly need exception lists. In addition, it is likely that stability rules for those properties would need to reflect IETF norms with arrangements for bringing the IETF and other communities into the discussion when tradeoffs are reviewed.

4.3. The need for exception lists

[[CREF5: Note in draft: this section is a partial placeholder and may need more elaboration.]]
Issues with exception lists and the requirements for them are discussed in Section 2 above and in RFC 5894 [RFC5894].

5. Proposed/ Alternative Changes to RFC 5892 for the issues first exposed by new code point U+08A1

NOTE IN DRAFT: See the comments in the Introduction, Section 1 and the first paragraph of each Subsection below for the status of the Subsections that follow. Each one, in combination with the material in Section 3 above, also provides information about the reasons why that particular strategy might or might not be appropriate.

When the term "Category" followed by an upper-case letter appears below, it is s reference to a rule in RFC 5892.

5.1. Disallow This New Code Point

This option is almost certainly too Arabic-specific and does not solve, or even address, the underlying problem. It also does not inherently generalize to non-decomposing precomposed code points that might be added in the future (whether to Arabic or other scripts)

even though one could add more code points to Category F in the same way.

If chosen by the community, this subsection would update the portion of the IDNA2008 specification that identifies rules for what characters are permitted [RFC5892] to disallow that code point.

With the publication of this document, Section 2.6 ("Exceptions (F)") of RFC 5892 [RFC5892] is updated by adding 08A1 to the rule in Category F so that the rule itself reads:

```
F: cp is in {00B7, 00DF, 0375, 03C2, 05F3, 05F4, 0640, 0660,
             0661, 0662, 0663, 0664, 0665, 0666, 0667, 0668,
             0669, 06F0, 06F1, 06F2, 06F3, 06F4, 06F5, 06F6,
             06F7, 06F8, 06F9, 06FD, 06FE, 07FA, 08A1, 0F0B,
             3007, 302E, 302F, 3031, 3032, 3033, 3034, 3035,
             303B, 30FB}
```

and then add to the subtable designated
"DISALLOWED -- Would otherwise have been PVALID"
after the line that begins "07FA", the additional line:

```
08A1; DISALLOWED # ARABIC LETTER BEH WITH HAMZA ABOVE
```

This has the effect of making the cited code point DISALLOWED independent of application of the rest of the IDNA rule set to the current version of Unicode. Those wishing to create domain name labels containing Beh with Hamza Above may continue to use the sequence

U+0628, ARABIC LETTER BEH
followed by

U+0654, ARABIC HAMZA ABOVE

which was valid for IDNA purposes in Unicode 5.0 and earlier and which continues to be valid.

In principle, much the same thing could be accomplished by using the IDNA "BackwardCompatible" category (IDNA Category G, RFC 5892 Section 5.3). However, that category is described as applying only when "property values in versions of Unicode after 5.2 have changed in such a way that the derived property value would no longer be PVALID or DISALLOWED". Because U+08A1 is a newly-added code point in Unicode 7.0.0 and no property values of code points in prior versions have changed, category G does not apply. If that section of RFC 5892 were to be replaced in the future, perhaps consideration should be

given to adding Normalization Stability and other issues to that description but, at present, it is not relevant.

5.2. Disallow This New Code Point and All Future Precomposed Additions that Do Not Decompose

At least in principle, the approach suggested above (Section 5.1) could be expanded to disallow all future allocations of non-decomposing precomposed characters. This would probably require either a new Unicode property to identify such characters and/or more emphasis on the manual, individual code point, checking of the new Unicode version review proces (i.e,. not just application of the existing rules and algorithm). It might require either a new rule in IDNA or a modification to the structure of Category F to make additions less tedious. It would do nothing for different ways to form identical characters within the same script that were not associated with decomposition and so would have to be used in conjunction with other appropaches. Finally, for scripts (such as Arabic) where there is a very strong preference to avoid combining sequences, this approach would exclude exactly the wrong set of characters.

5.3. Disallow the combining sequences for these characters

As in the approach discussed in Section 5.1, this approach is too Arabic-specific to address the more general problem. However, it illustrates a single-script approach and a possible mechanism for excluding combining sequences whose handling is connected to language information (information that, as discussed above, is not relevant to the DNS).

If chosen by the community, this subsection would update the portion of the IDNA2008 specification that identifies contextual rules [RFC5892] to prohibit (combining) Hamza Above (U+0654) in conjunction with Arabic BEH (U+0628), HAH (U+062D), and REH (U+0631). Note that the choice of this option is consistent with the general preference for precomposed characters discussed above but would ban some labels that are valid today and that might, in principle, be in use.

The required prohibition could be imposed by creating a new contextual rule in RFC 5892 to constrain combining sequences containing Hamza Above.

As the Unicode Standard points out at some length [Unicode70-Arabic], Hamza is a problematic abstract character and the "Hamza Above" construction even more so. IDNA has historically associated characters whose use is reasonable in some contexts but not others with the special derived property "CONTEXTO" and then specified

specific, context-dependent, rules about where they may be used. Because Hamza Above is problematic (and spawns edge cases, as discussed in the Unicode Standard section cited above), it was suggested that a contextual rule might be appropriate. There are at least two reasons why a contextual rule would not be suitable for the present situation.

1. As discussed above, the present situation is a normalization stability and predictability problem, not a contextual one. Had the same issues arisen with a newly-added precomposed character that could previously be constructed from non-problematic base and combining characters, it would be even more clearly a normalization issue and, following the principles discussed there and particularly in UAX 15 [UAX15-Exclusion], might not have been assigned at all.

2. The contextual rule sets are designed around restricting the use of code points to a particular script or adjacent to particular characters within that script. Neither of these cases applies to the newly-added character even if one could imagine rules for the use of Hamza Above (U+0654) that would reflect the considerations of Chapter 8 of Unicode 6.2. Even had the latter been desired, it would be somewhat late now -- Hamza Above has been present as a combining character (U+0654) in many versions of Unicode. While that section of the Unicode Standard describes the issues, it does not provide actionable guidance about what to do about it for cases going forward or when visual identity is important.

5.4. Use Combinnig Classes to Develop Additional Contextual Rules

This option may not be of any practical use, but Unicode supports a property called "Combining_Class". That property has been used in IDNA only to construct a contextual rule for Zero-Width Non-Joiner [RFC5892, Appendix A.1] but speculation has arisen during discussions of work on Arabic combining characters and rendering [UTR53] as to whether Combining Classes could be used to build additional contextual rules that would restrict problematic cases. Unless such rules were applied only to new code points, they would also not be backward compatable.

The question of whether Combining Classes could be used to reduce the number of problematic labels is at least worth examination.

5.5. Disallow all Combining Characters for Specific Scripts

[[CREF6: This subsection needs to be turned into prose, but the follow bullet points are probably sufficient to identify the issues.]]

- Might work for Arabic and other "precomposed preference" scripts if those can be identified in an orderly and stable way (see Section 3.3.4.1; recommended by the Arabic language community for IDNs [RFC5564]).

- Unworkable for Latin because many characters that do not decompose are, at least in part, historical accidents resulting from combining prior national standards (this probably may exist for other scripts as well).

- No effect at all on special-use representations of identical characters within a script (see Section 3.3.2.1 and Section 3.3.2.2).

- Not backwards compatible.

5.6. Do Nothing Other Than Warn

A recommendation from UTC and others has been to simply warn registries, at all levels of the tree, to be careful with this set of characters. Doing that well would probably require making language distinctions within zones, which would violate the important IDNA principles that labels are not necessarily "words", do not carry language information, and may, at the protocol level, even deliberately mix languages and scripts. It is also problematic because the relevant set of characters is not easily defined in a precise way. This suggestion is problematic because the DNS and IDNA cannot make or enforce language distinctions, but it would avoid having the IETF either invalidate label strings that are potentially now in use or creating inconsistencies among the characters that combine with selected base characters but that also have precomposed forms that do not have decompositions. The potential would still exist for registries to respect the warning and deprecate such labels if they existed.

More generally, while there are already requirements in IDNA for registries to be knowledgeable and responsible about the labels they register (a separate document discusses that requirement [Klensin-rfc5891bis]), experience indicates that those requirements are often ignored. At least as important, warning registries about what should or should not be registered and even calling out specific code points as dangerous and in need of extra attention [Freytag-dangerous] does nothing to address the many cases in which lookup-time checking for IDNA conformance and deliberately misleading label constructions is important.

5.7. Normalization Form IETF (NFI))

The most radical possibility for the comparison issue would be to decide that none of the Unicode Normalization Forms specified in UAX 15 [UAX15] are adequate for use with the DNS because, contrary to their apparent descriptions, normalization tables are actually determined using language information. However, use of language information is unacceptable for IDNA for reasons described elsewhere in this document. The remedy would be to define an IETF-specific (or DNS-specific) normalization form (sometimes called "NFI" in discussions), building on NFC but adhering strictly to the rule that normalization causes two different forms of the same character (glyph image) within the same script to be treated as equal. In practice such a form could be implemented for IDNA purposes as an additional rule within RFC 5892 (and its successors) that constituted an exception list for the NFC tables. For this set of characters, the special IETF normalization form would be equivalent to the exclusion discussed in Section 5.3 above.

An Internet-identifier-specific normalization form, especially if specified somewhat separately from the IDNA core, would have a small marginal advantage over the other strategies in this section (or in combination with some of them), even though most of the end result and much of the implementation would be the same in practice. While the design of IDNA requires that strings be normalized as part of the process of determining label validity (and hence before either storage of values in the DNS or name resolution), there is an ongoing debate about whether normalization should be performed before storing a string or putting it on the wire or only when the string is actually compared or otherwise used.

If a normalization procedure with the right properties for the IETF was defined, that argument could be bypassed and the best decisions made for different circumstances. The separation would also allow better comparison of strings that lack language context in applications environments in which the additional processing and character classifications of IDNA and/or PRECIS were not applicable. Having such a normalization procedure defined outside IDNA would also minimize changes to IDNA itself, which is probably an advantage.

If the new normalizstion form were, in practice, simply an overlay on NFC with modifications dictated by exception and/or property lists, keeping its definition separate from IDNA would also avoid interweaving those exceptions and property lists with the rules and categories of IDNA itself, avoiding some unnecessary complexity.

6. Editorial clarification to RFC 5892

Verified RFC Editor Erratum 3312 [RFC5892Erratum] provides a clarification to Appendix A and Section A.1 of RFC 5892. This section of this document updates the RFC to apply that clarification.

1. In Appendix A, add a new paragraph after the paragraph that begins "The code point...". The new paragraph should read:

   "For the rule to be evaluated to True for the label, it MUST be evaluated separately for every occurrence of the Code point in the label; each of those evaluations must result in True."

2. In Appendix A, Section A.1, replace the "Rule Set" by

```
Rule Set:
  False;
  If Canonical_Combining_Class(Before(cp)) .eq.  Virama Then True;
  If cp .eq. \u200C And
         RegExpMatch((Joining_Type:{L,D})(Joining_Type:T)*cp
    (Joining_Type:T)*(Joining_Type:{R,D})) Then True;
```

7. Acknowledgements

The Unicode 7.0.0 changes were extensively discussed within the IAB's Internationalization Program. The authors are grateful for the discussions and feedback there, especially from Andrew Sullivan and David Thaler. Additional information was requested and received from Mark Davis and Ken Whistler and while they probably do not agree with the necessity of excluding this code point or taking even more drastic action as their responsibility is to look at the Unicode Consortium requirements for stability, the decision would not have been possible without their input. Thanks to Bill McQuillan and Ted Hardie for reading versions of the document carefully enough to identify and report some confusing typographical errors. Several experts and reviewers who prefer to remain anonymous also provided helpful input and comments on preliminary versions of this document.

8. IANA Considerations

When the IANA registry and tables are updated to reflect Unicode 7.0.0, changes should be made according to the decisions the IETF makes about Section 5.

9. Security Considerations

From at least one point of view, this document is entirely a discussion of a security issue or set of such issues. While the "similar-looking characters" issue that has been a concern since the earliest days of IDNs [HomographAttack] and that has driven assorted "character confusion" projects [ICANN-VIP], if a user types in a string on one device and can get different results that do not compare equal when it is typed on a different device (with both behaving correctly and both keyboards appearing to be the same and for the same script) then all security mechanism that depend on the underlying identifiers, including the practical applications of DNS response integrity checks via DNSSEC [RFC4033] and DNS-embedded public key mechanisms [RFC6698], are at risk if different parties, at least one of them malicious, obtain or register some of the identical-appearing and identically-typed strings and get them into appropriate zones.

Mechanisms that depend on trusting registration systems (e.g., registries and registrars in the DNS IDN case, see Section 5.6 above) are likely to be of only limited utility because fully-qualified domains that may be perfectly reasonable at the first level or two of the DNS may have differences of this type deep in the tree, into levels where name management, and often accountability, are weak. Similar issues obviously apply when names are user-selected or unmanaged.

When the issue is not a deliberate attack but simple accidental confusion among similar strings, most of our strategies depend on the acceptability of false negatives on matching if there is low risk of false positives (see, for example, the discussion of false negatives in identifier comparison in Section 2.1 of RFC 6943 [RFC6943]). Aspects of that issue appear in, for example, RFC 3986 [RFC3986] and the PRECIS effort [RFC8264]. However, because the cases covered here are connected, not just to what the user sees but to what is typed and where, there is an increased risk of false positives (accidental as well as deliberate).

[[CREF7: Note in Draft: The paragraph that follows was written for a much earlier version of this document. It is obsolete, but is being retained as a placeholder for future developments.]]

This specification excludes a code point for which the Unicode-specified normalization behavior could result in two ways to form a visually-identical character within the same script not comparing equal. That behavior could create a dream case for someone intending to confuse the user by use of a domain name that looked identical to

another one, was entirely in the same script, but was still considered different.

Internet Security in areas that involve internationalized identifiers that might contain the relevant characters is therefore significantly dependent on some effective resolution for the issues identified in this document, not just hand waving, devout wishes, or appointment of study committees about it.

10. References

10.1. Normative References

[RFC5137] Klensin, J., "ASCII Escaping of Unicode Characters", BCP 137, RFC 5137, DOI 10.17487/RFC5137, February 2008, <https://www.rfc-editor.org/info/rfc5137>.

[RFC5890] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework", RFC 5890, DOI 10.17487/RFC5890, August 2010, <https://www.rfc-editor.org/info/rfc5890>.

[RFC5892] Faltstrom, P., Ed., "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)", RFC 5892, DOI 10.17487/RFC5892, August 2010, <https://www.rfc-editor.org/info/rfc5892>.

[RFC5892Erratum] "RFC5892, "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)", August 2010, Errata ID: 3312", Errata ID 3312, August 2012, <http://www.rfc-editor.org/errata_search.php?rfc=5892>.

[RFC5894] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Background, Explanation, and Rationale", RFC 5894, DOI 10.17487/RFC5894, August 2010, <https://www.rfc-editor.org/info/rfc5894>.

[RFC6943] Thaler, D., Ed., "Issues in Identifier Comparison for Security Purposes", RFC 6943, DOI 10.17487/RFC6943, May 2013, <https://www.rfc-editor.org/info/rfc6943>.

[RFC8264] Saint-Andre, P. and M. Blanchet, "PRECIS Framework: Preparation, Enforcement, and Comparison of Internationalized Strings in Application Protocols", RFC 8264, DOI 10.17487/RFC8264, October 2017, <https://www.rfc-editor.org/info/rfc8264>.

[UAX15] Davis, M., Ed., "Unicode Standard Annex #15: Unicode Normalization Forms", June 2014, <http://www.unicode.org/reports/tr15/>.

[UAX15-Exclusion]
"Unicode Standard Annex #15: ob. cit., Section 5", <http://www.unicode.org/reports/tr15/#Primary_Exclusion_List_Table>.

[UAX15-Versioning]
"Unicode Standard Annex #15, ob. cit., Section 3", <http://www.unicode.org/reports/tr15/#Versioning>.

[Unicode5]
The Unicode Consortium, "The Unicode Standard, Version 5.0", ISBN 0-321-48091-0, 2007.

Boston, MA, USA: Addison-Wesley. ISBN 0-321-48091-0. This printed reference has now been updated online to reflect additional code points. For code points, the reference at the time RFC 5890-5894 were published is to Unicode 5.2.

[Unicode62]
The Unicode Consortium, "The Unicode Standard, Version 6.2.0", ISBN 978-1-936213-07-8, 2012, <http://www.unicode.org/versions/Unicode6.2.0/>.

Preferred citation: The Unicode Consortium. The Unicode Standard, Version 6.2.0, (Mountain View, CA: The Unicode Consortium, 2012. ISBN 978-1-936213-07-8)

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

Preferred Citation: The Unicode Consortium. The Unicode Standard, Version 7.0.0, (Mountain View, CA: The Unicode Consortium, 2014. ISBN 978-1-936213-09-2)

[Unicode70-Arabic]
"The Unicode Standard, Version 7.0.0, ob.cit., Chapter 9.2: Arabic", Chapter 9, 2014, <http://www.unicode.org/versions/Unicode7.0.0/ch09.pdf>.

Subsection titled "Encoding Principles", paragraph numbered 4, starting on page 362.

[Unicode70-CompatDecomp]
"The Unicode Standard, Version 7.0.0, ob.cit., Chapter 2.3: Compatibility Characters", Chapter 2, 2014, <http://www.unicode.org/versions/Unicode7.0.0/ch02.pdf>.

Subsection titled "Compatibility Decomposable Characters" starting on page 26.

[Unicode70-Design]
"The Unicode Standard, Version 7.0.0, ob.cit., Chapter 2.2: Unicode Design Principles", Chapter 2, 2014, <http://www.unicode.org/versions/Unicode7.0.0/ch02.pdf>.

[Unicode70-Hamza]
"The Unicode Standard, Version 7.0.0, ob.cit., Chapter 9.2: Arabic", Chapter 9, 2014, <http://www.unicode.org/versions/Unicode7.0.0/ch09.pdf>.

Subsection titled "Combining Hamza Above" starting on page 378.

[Unicode70-Overlay]
"The Unicode Standard, Version 7.0.0, ob.cit., Chapter 2.2: Unicode Design Principles", Chapter 2, 2014, <http://www.unicode.org/versions/Unicode7.0.0/ch02.pdf>.

Subsection titled "Non-decomposition of Overlaid Diacritics" starting on page 64.

[Unicode70-Stability]
"The Unicode Standard, Version 7.0.0, ob.cit., Chapter 2.2: Unicode Design Principles", Chapter 2, 2014, <http://www.unicode.org/versions/Unicode7.0.0/ch02.pdf>.

Subsection titled "Stability" starting on page 23 and containing a link to http://www.unicode.org/policies/stability_policy.html..

[UTS46] Davis, M. and M. Suignard, "Unicode Technical Standard #46: Unicode IDNA Compatibility Processing", Version 7.0.0, June 2014, <http://unicode.org/reports/tr46/>.

10.2. Informative References

[Dalby] Dalby, A., "Dictionary of Languages: The definitive reference to more than 400 languages", Columbia Univeristy Press , 2004.

pages 206-207

[Daniels] Daniels, P. and W. Bright, "The World's Writing Systems", Oxford University Press , 1986.

page 744

[Freytag-dangerous]
Freytag, A., Klensin, J., and A. Sullivan, "Those Troublesome Characters: A Registry of Unicode Code Points Needing Special Consideration When Used in Network Identifiers", June 2017, <https://datatracker.ietf.org/doc/draft-freytag-troublesome-characters/>.

[HomographAttack]
Gabrilovich, E. and A. Gontmakher, "The Homograph Attack", Communications of the ACM 45(2):128, February 2002, <http://www.cs.technion.ac.il/~gabr/papers/homograph_full.pdf>.

[ICANN-VIP]
ICANN, "The IDN Variant Issues Project: A Study of Issues Related to the Management of IDN Variant TLDs (Integrated Issues Report)", February 2012, <https://www.icann.org/en/system/files/files/idn-vip-integrated-issues-final-clean-20feb12-en.pdf>.

[Klensin-rfc5891bis]
Klensin, J., "Internationalized Domain Names in Applications (IDNA): Registry Restrictions and Recommendations", September 2017, <https://datatracker.ietf.org/doc/draft-klensin-idna-rfc5891bis/>.

[Omniglot-Fula]
Ager, S., "Omniglot: Fula (Fulfulde, Pulaar, Pular'Fulaare)", <http://www.omniglot.com/writing/fula.htm>.

Captured 2015-01-07

[RFC0020] Cerf, V., "ASCII format for network interchange", STD 80, RFC 20, DOI 10.17487/RFC0020, October 1969, <https://www.rfc-editor.org/info/rfc20>.

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

[RFC3492] Costello, A., "Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA)", RFC 3492, DOI 10.17487/RFC3492, March 2003, <https://www.rfc-editor.org/info/rfc3492>.

[RFC3986] Berners-Lee, T., Fielding, R., and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax", STD 66, RFC 3986, DOI 10.17487/RFC3986, January 2005, <https://www.rfc-editor.org/info/rfc3986>.

[RFC4033] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "DNS Security Introduction and Requirements", RFC 4033, DOI 10.17487/RFC4033, March 2005, <https://www.rfc-editor.org/info/rfc4033>.

[RFC5564] El-Sherbiny, A., Farah, M., Oueichek, I., and A. Al-Zoman, "Linguistic Guidelines for the Use of the Arabic Language in Internet Domains", RFC 5564, DOI 10.17487/RFC5564, February 2010, <https://www.rfc-editor.org/info/rfc5564>.

[RFC6452] Faltstrom, P., Ed. and P. Hoffman, Ed., "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA) - Unicode 6.0", RFC 6452, DOI 10.17487/RFC6452, November 2011, <https://www.rfc-editor.org/info/rfc6452>.

[RFC6698] Hoffman, P. and J. Schlyter, "The DNS-Based Authentication of Named Entities (DANE) Transport Layer Security (TLS) Protocol: TLSA", RFC 6698, DOI 10.17487/RFC6698, August 2012, <https://www.rfc-editor.org/info/rfc6698>.

[Unicode32]
The Unicode Consortium, "The Unicode Standard, Version 3.2.0".

The Unicode Standard, Version 3.2.0 is defined by The Unicode Standard, Version 3.0 (Reading, MA, Addison-Wesley, 2000. ISBN 0-201-61633-5), as amended by the Unicode Standard Annex #27: Unicode 3.1 (http://www.unicode.org/reports/tr27/) and by the Unicode Standard Annex #28: Unicode 3.2 (http://www.unicode.org/reports/tr28/).

[UTR53] Unicode Consortium, "Proposed Draft: Unicode Technical Report #53: Unicode Arabic Mark Ordering Algorithm", August 2017, <http://www.unicode.org/reports/tr53/>.

Note: this is a Proposed Draft, out for public review when this version of the current I-D is posted, and should not be considered either an approved/ final document or a stable reference.

Appendix A. Change Log

RFC Editor: Please remove this appendix before publication.

A.1. Changes from version -00 (2014-07-21)to -01

- Version 01 of this document is an extensive rewrite and reorganization, reflecting discussions with UTC members and adding three more options for discussion to the original proposal to simply disallow the new code point.

A.2. Changes from version -01 (2014-12-07) to -02

Corrected a typographical error in which Hamza Above was incorrectly listed with the wrong code point.

A.3. Changes from version -02 (2014-12-07) to -03

Corrected a typographical error in the Abstract in which RFC 5892 was incorrectly shown as 5982.

A.4. Changes from version -03 (2015-01-06) to -04

- Explicitly identified the applicability of U+08A1 with Fula and added references that discuss that language and how it is written.

- Updated several Unicode 6.2 references to point to Unicode 7.0 since the latter is now available in stable form (it was done when work on this I-D started).

- Extensively revised to discuss the non-Arabic cases, non-decomposing diacritics, other types of characters that don’t compare equal after normalization, and more general problem and approaches.

A.5. Changes from version -04 (2015-03-11) to -05

- Modified a few citation labels to make them more obvious.
- Restructured Section 1 and added additional terminology comments.
- Added discussion about non-decomposable character cases, including the "slash" example, and associated references for which -04 contained only placeholders.
- The examples and discussion of Latin script issues has been expanded considerably. It is unfortunate that many readers in the IETF community apparently cannot understand examples well enough to believe a problem is significant unless they is a discussion of Latin script examples, but, at least for this working draft, that is the way it is.
- Rewrote the discussion of several of the alternatives and added the discussion of combining classes.
- Rewrote and extended the discussion of the "warn only" alternative.
- Several other sections modified to improve technical or editorial clarity.
- Note that, while some references have been updated, others have not. In particular, Unicode references are still tied to versions 6 or 7. In some cases, those non-historical references are and will remain appropriate; others will best be replaced with information about current versions of documents.

Authors' Addresses

John C Klensin
1770 Massachusetts Ave, Ste 322
Cambridge, MA 02140
USA

Phone: +1 617 245 1457
Email: john-ietf@jck.com

Patrik Faltstrom
Netnod
Franzengatan 5
Stockholm 112 51
Sweden

Phone: +46 70 6059051
Email: paf@netnod.se

Network Working Group
Internet-Draft
Updates: 5890, 5891, 5894 (if approved)
Intended status: Standards Track
Expires: March 16, 2018

J. Klensin

A. Freytag
ASMUS, Inc.

September 12, 2017

Internationalized Domain Names in Applications (IDNA): Registry Restrictions and Recommendations
draft-klensin-idna-rfc5891bis-01

Abstract

The IDNA specifications for internationalized domain names combine rules that determine the labels that are allowed in the DNS without violating the protocol itself and an assignment of responsibility, consistent with earlier specifications, for determining the labels that are allowed in particular zones. Conformance to IDNA by registries and other implementations requires both parts. Experience strongly suggests that the language describing those responsibilities was insufficiently clear to promote safe and interoperable use of the specifications and that more details and some specific examples would have been helpful. Without making any substantive changes to IDNA, this specification updates two of the core IDNA documents (RFC 5980 and 5891) and the IDNA explanatory document (RFC 5894) to provide that guidance and to correct some technical errors in the descriptions.

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 https://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 March 16, 2018.

Copyright Notice

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction

Parts of the specifications for Internationalized Domain Names in Applications (IDNA) [RFC5890] [RFC5891] [RFC5894] (collectively known, along with RFC 5892 [RFC5892], RFC 5893 [RFC5893] and updates to them, as "IDNA2008" (or just "IDNA") impose a requirement that domain name system (DNS) registries restrict the characters they allow in domain name labels (see Section 2 below), and the contents and structure of those labels. That requirement and restriction are consistent with the "trustee for the community" requirements of the original specification for DNS naming and authority [RFC1591]. The restrictions are intended to limit the permitted characters and strings to those for which the registries or their advisers have a

thorough understanding and for which they are willing to take responsibility.

That provision is centrally important because it recognized that historical relationships and variations among scripts and writing systems, the continuing evolution of those systems, differences in the uses of characters among languages (and locations) that use the same script, and so on make it impossible for a single list of characters and simple rules to be able to generate an "if we use these, we will be safe from confusion and various attacks" guideline.

Instead, the algorithm and rules of RFC 5981 and 5982 eliminate many of the most dangerous and otherwise problematic cases, but cannot eliminate the need for registries and registrars to understand what they are doing and taking responsibility for the decisions they make.

The way in which the IDNA2008 specifications expressed these requirements may have obscured the intention that they actually are requirements. Section 2.3.2.3 of the Definitions document [RFC5890] mentions the need for the restrictions, indicates that they are mandatory, and points the reader to section 4.3 of the Protocol document [RFC5891], which in turn points to Section 3.2 of the Rationale document [RFC5894], with each document providing further detail, discussion, and clarification.

This specification is intended to unify and clarify these requirements for registry decisions and responsibility and to emphasize the importance of registry restrictions at all levels of the DNS. It also makes a specific recommendation for character repertoire subsetting intermediate between the code points allowed by RFC 5891 and 5892 and those allowed by individual registries. It does not alter the basic IDNA2008 protocols and rules themselves in any way.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Registry Restrictions in IDNA2008

As mentioned above, IDNA2008 specifies that the registries for each zone in the DNS that supports IDN labels are required to develop and apply their own rules to restrict the allowable labels, including limiting characters they allow to be used in labels in that zone. The chosen list MUST BE smaller than the collection of code points specified as "PVALID", "CONTEXTJ", and "CONTEXTO" by the rules established by the protocols themselves. The latter two categories, and labels containing any characters that are normally part of a

script written right to left [RFC5893], require that additional rules, specified in the protocols and known as "contextual rules" and "bidi rules", be applied. The entire collection of rules and restrictions required by the IDNA2008 protocols themselves are known as "protocol restrictions".

As mentioned above, registries may apply (and generally are required to apply) additional rules to further restrict the list of permitted code points, contextual rules (perhaps applied to normally PVALID code points) that apply additional restrictions, and/or restrictions on labels. The most obvious of those restrictions include provisions for restricting suggested new registrations based on conflicts with labels already registered in the zone and specifications of what constitutes such conflicts based on the properties of the labels in question. They further include prohibitions on code points and labels that are not consistent with the intended function of the zone or the subtree in which it is embedded (see Section 3) or limitations on where in a label allowable code points may be placed.

These per-registry (or per-zone) rules are commonly known as "registry restrictions" to distinguish them from the protocol restrictions described above. By necessity, the latter are somewhat generic, having to cater both to the union of the needs for all zones, as well as to the most permissive zones. In consequence, additional Registry restrictions are essential to provide for the necessary security in the face of the tremendous variations and differences in writing systems, their ongoing evolution and development, as well as the human ability to recognize and distinguish characters in different scripts around the world and under different circumstances.

3. Progressive Subsets of Allowed Characters

The algorithm and rules of RFC 5891 and 5892 set an absolute upper bound on the code points that can be used in domain name labels; registries MUST NOT include code points unless they are allowed by those rules. Each registry that intends to allow IDN registrations MUST then determine which code points will be allowed by that registry. It SHOULD also consider additional rules, including contextual and whole label restrictions that provide further protection for registrants and users. For example, the widely-used principle that bars labels containing characters from more than one script is not an IDNA2008 requirement. It has been adopted by many registries but, as Section 4.4 of RFC 5890 indicates, there may be circumstances in which is it not required or appropriate.

In formulating their own rules, registries SHOULD normally consult carefully-developed consensus recommendations about global maximum

repertoires to be used such as the ICANN Maximal Starting Repertoire 2 (MSR-2) for the Development of Label Generation Rules for the Root Zone [ICANN-MSR2] (or its successor documents). Additional recommendations of similar quality about particular scripts or languages exist, including, but not limited to, the RFCs for Cyrillic [RFC5992] or Arabic Language [RFC5564] or script-based repertoires from the approved ICANN Root Zone Label Generation Rules (LGR-1) [ICANN-LGR1] (or its successor documents).

It is the responsibility of the registry to determine which, if any, of those recommendations are applicable and to further subset or extend them as needed. For example, several of the recommendations are designed for the root zone and therefore exclude digits and U+002D HYPHEN-MINUS; this restriction is not generally appropriate for other zones. On the other hand, some zones may be designed to not cater for all users of a given script, but perhaps only for the needs of selected languages, in which case a more selective repertoire may be appropriate.

In making these determinations, a registry SHOULD follow the IAB guidance in RFC 6912 [RFC6912]. Those guidelines include a number of principles for use in making decisions about allowable code points. In addition, that document notes that the closer a particular zone is to the root, the more restrictive the space of permitted labels should be. RFC 5894 provides some suggestions for any registry that may decide to reduce opportunities for confusion or attacks by constructing policies that disallow characters used in historic writing systems (whether these be archaic scripts or extensions of modern scripts for historic or obsolete orthographies) or characters whose use is restricted to specialized, or highly technical contexts. These suggestions were among the principles guiding the design of ICANN’s Maximal Starting Repertoires [LGR-Procedure].

Particularly for a zone for which all labels to be delegated are not for the use of the same organization or enterprise, a registry decision to allow only those code points in the full repertoire of the MSR (plus digits and hyphen) would already avoid a number of issues inherent in a more permissive policy like "use anything permitted by IDNA2008", while still supporting the native languages and scripts for the vast majority of users today. However, it is unlikely, by itself, to fully satisfy the mandate set out above for three reasons.

1. The MSR, like the set of code points permissible under IDNA2008 itself, was conceived merely as an upper bound on permissible letter code points (it excludes digits and the hyphen). It was always intended to be used as a starting point for setting registry policy, with the expectation that some of the code

points in the MSR would not be included in the final registry policy, whether for lack of actual usage, or for being inherently problematic.

2. It was recognized that many scripts require contextual rules for many more code points than are covered by CONTEXTO or CONTEXTJ rules defined in IDNA2008. This is particularly true for combining marks, typically used to encode diacritics, tone marks, vowel signs and the like. While, theoretically, any combining mark may occur in any context in Unicode, in practice rendering and other software that users rely on in viewing or entering labels will not support arbitrary combining sequences, or indeed arbitrary combinations of code points, in the case of complex scripts.

   Contextual rules are required to limit allowable code point sequences to those that can be expected to be rendered reliably. Identifying those requires knowledge about the way code points are used in a script, whence the mandate for registries to only support code points they understand. In this, some of the other recommendations, such as the Informational RFCs for specific scripts (e.g., Cyrillic [RFC5992]) or languages (e.g., Arabic [RFC5564] or Chinese [RFC4713]), or the Root Zone LGRs developed by ICANN, may provide useful guidance.

3. Third, because of the widely accepted practice of limiting any given label to a single script, a universal repertoire, such as the MSR, would have to be divided on a per script basis into subrepertoires to make it useful, with some of those repertoires overlapping, for example, in the case of East Asian shared usage of the Han ideographs.

Registries choosing to make exceptions and allow code points that recommendations such as the MSR do not allow should make such decisions only with great care and only if they have considerable understanding of, and great confidence in, their appropriateness. The obvious exception from the MSR would be to allow digits and the hyphen. Neither were allowed by the MSR, but only because they are not allowed in the Root Zone.

Nothing in this document permits a registry to allow code points or labels that are disallowed or otherwise prohibited by IDNA2008.

4. Other corrections and updates

After the initial IDNA2008 documents were published (and RFC 5892 was updated for Unicode 6.0 by RFC 6452 [RFC6452]) several errors or instances of confusing text were noted. For the convenience of the

community, the relevant corrections for RFC 5890 and 5891 are noted below and update the corresponding documents. There are no errata for RFC 5893 or 5894 as of the date this document was published. Because further updates to RFC 5892 would require addressing other pending issues, the outstanding erratum for that document is not considered here. For consistency with the original documents, references to Unicode 5.0 are preserved.

4.1. Updates to RFC 5890

The outstanding errata against RFC 5890 (Errata ID 4695, 4696, 4823, and 4824 [RFC-Editor-5890Errata]) are all associated with the same issue, the number of Unicode characters that can be associated with a maximum-length (63 octet) A-label. In retrospect and contrary to some of the suggestions in the errata, that value should not be expressed in octets because RFC 5890 and the other IDNA 2008 documents are otherwise careful to not specify Unicode encoding forms but, instead, work exclusively with Unicode code points. Consequently the relevant material in RFC 5890 should be corrected as follows:

Section 2.3.2.1

Old: expansion of the A-label form to a U-label may produce strings that are much longer than the normal 63 octet DNS limit (potentially up to 252 characters).

New: expansion of the A-label form to a U-label may produce strings that are much longer than the normal 63 octet DNS limit (See Section 4.2).

Comment: If the length limit is going to be a source of confusion or careful calculations, it should appear in only one place.

Section 4.2

Old: Because A-labels (the form actually used in the DNS) are potentially much more compressed than UTF-8 (and UTF-8 is, in general, more compressed that UTF-16 or UTF-32), U-labels that obey all of the relevant symmetry (and other) constraints of these documents may be quite a bit longer, potentially up to 252 characters (Unicode code points).

New: A-labels (the form actually used in the DNS) and the Punycode algorithm used as part of the process to produce them [RFC3492] are strings that are potentially much more compressed than any standard Unicode Encoding Form. [[CREF1: Do we need a reference for this here??]] A 63 octet A-label cannot

represent more than 58 Unicode code points (four octet overhead and the requirement that at least one character lie outside the ASCII range) but implementations allocating buffer space for the conversion should allow significantly more space depending on the encoding form they are using.

4.2. Updates to RFC 5891

Errata ID 3969: Improve reference for combining marks There is only one erratum for RFC 5891, Errata ID 3969 [RFC5891Erratum]. Combining marks are explained in the cited section, but not, as the text indicates, exactly defined.

Old: The Unicode string MUST NOT begin with a combining mark or combining character (see The Unicode Standard, Section 2.11 [Unicode] for an exact definition).

New: The Unicode string MUST NOT begin with a combining mark or combining character (see The Unicode Standard, Section 2.11 [Unicode] for an explanation and Section 3.6, definition D52) for an exact definition).

Comment: When RFC 5891 is actually updated, the references in the text should be updated to the current version of Unicode and the section numbers checked.

5. Related Discussions

This document is one of a series of measures that have been suggested to address IDNA issues raised in other documents, including mechanisms for dealing with combining sequences and single-code point characters with the same appearance that normalization neither combines nor decomposes as IDNA2008 assumed [IDNA-Unicode], including the IAB response to that issue [IAB-2015], and to take a higher-level view of issues, demands, and proposals for new uses of the DNS. Those documents also include a discussion of issues with IDNA and character graphemes for which abstractions exist in Unicode in precomposed form but that can be generated from combining sequences and a suggested registry of code points known to be problematic [Freytag-troublesome]. The discussion of combining sequences and non-decomposing characters is intended to lay the foundation for an actual update to the IDNA code points document [RFC5892]. Such an update will presumably also address the existing errata against that document.

6. Security Considerations

As discussed in IAB recommendations about internationalized domain names [RFC4690], [RFC6912], and elsewhere, poor choices of strings for DNS labels can lead to opportunities for attacks, user confusion, and other issues less directly related to security. This document clarifies the importance of registries carefully establishing design policies for the labels they will allow and that having such policies and taking responsibility for them is a requirement, not an option. If that clarification is useful in practice, the result should be an improvement in security.

7. Acknowledgments

Many thanks to Patrik Faltstrom who provided an important review on the initial version.

8. IANA Considerations

[[CREF2: RFC Editor: Please remove this section before publication.]]

This memo includes no requests to or actions for IANA. In particular, it does not contain any provisions that would alter any IDNA-related registries or tables.

9. References

9.1. Normative References

[ICANN-LGR1] ICANN, "Root Zone Label Generation Rules (LGR-1)", June 2015, <https://www.icann.org/resources/pages/root-zone-lgr-2015-06-21-en>.

[ICANN-MSR2] ICANN, "Maximal Starting Repertoire Version 2 (MSR-2) for the Development of Label Generation Rules for the Root Zone", April 2015, <https://www.icann.org/news/announcement-2-2015-04-27-en>.

[RFC1591] Postel, J., "Domain Name System Structure and Delegation", RFC 1591, DOI 10.17487/RFC1591, March 1994, <https://www.rfc-editor.org/info/rfc1591>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC5890] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework", RFC 5890, DOI 10.17487/RFC5890, August 2010, <https://www.rfc-editor.org/info/rfc5890>.

[RFC5891] Klensin, J., "Internationalized Domain Names in Applications (IDNA): Protocol", RFC 5891, DOI 10.17487/RFC5891, August 2010, <https://www.rfc-editor.org/info/rfc5891>.

[RFC5891Erratum] "RFC 5891, "Internationalized Domain Names in Applications (IDNA): Protocol"", Errata ID 3969, April 2014, <http://www.rfc-editor.org/errata_search.php?rfc=5891>.

[RFC5894] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Background, Explanation, and Rationale", RFC 5894, DOI 10.17487/RFC5894, August 2010, <https://www.rfc-editor.org/info/rfc5894>.

9.2. Informative References

[Freytag-troublesome] Freytag, A., Klensin, J., and A. Sullivan, "Those Troublesome Characters: A Registry of Unicode Code Points Needing Special Consideration When Used in Network Identifiers", June 2017, <draft-freytag-troublesome-characters-01>.

[IAB-2015] Internet Architecture Board (IAB), "IAB Statement on Identifiers and Unicode 7.0.0", February 2015, <https://www.iab.org/documents/correspondence-reports-documents/2015-2/iab-statement-on-identifiers-and-unicode-7-0-0/>.

[IDNA-Unicode] Klensin, J. and P. Falstrom, "IDNA Update for Unicode 7.0.0", September 2017, <draft-klensin-idna-5892upd-unicode70-05>.

[LGR-Procedure] Internet Corporation for Assigned Names and Numbers (ICANN), "Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels", March 2013, <https://www.icann.org/en/system/files/files/draft-lgr-procedure-20mar13-en.pdf>.

[RFC-Editor-5890Errata]
RFC Editor, "RFC Errata: RFC 5890, "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework", August 2010", Note to RFC Editor: Please figure out how you would like this referenced and make it so., Captured 2017-09-10, 2016, <https://www.rfc-editor.org/errata_search.php?rfc=5890>.

[RFC3492] Costello, A., "Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA)", RFC 3492, DOI 10.17487/RFC3492, March 2003, <https://www.rfc-editor.org/info/rfc3492>.

[RFC4690] Klensin, J., Faltstrom, P., Karp, C., and IAB, "Review and Recommendations for Internationalized Domain Names (IDNs)", RFC 4690, DOI 10.17487/RFC4690, September 2006, <https://www.rfc-editor.org/info/rfc4690>.

[RFC4713] Lee, X., Mao, W., Chen, E., Hsu, N., and J. Klensin, "Registration and Administration Recommendations for Chinese Domain Names", RFC 4713, DOI 10.17487/RFC4713, October 2006, <https://www.rfc-editor.org/info/rfc4713>.

[RFC5564] El-Sherbiny, A., Farah, M., Oueichek, I., and A. Al-Zoman, "Linguistic Guidelines for the Use of the Arabic Language in Internet Domains", RFC 5564, DOI 10.17487/RFC5564, February 2010, <https://www.rfc-editor.org/info/rfc5564>.

[RFC5892] Faltstrom, P., Ed., "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)", RFC 5892, DOI 10.17487/RFC5892, August 2010, <https://www.rfc-editor.org/info/rfc5892>.

[RFC5893] Alvestrand, H., Ed. and C. Karp, "Right-to-Left Scripts for Internationalized Domain Names for Applications (IDNA)", RFC 5893, DOI 10.17487/RFC5893, August 2010, <https://www.rfc-editor.org/info/rfc5893>.

[RFC5992] Sharikov, S., Miloshevic, D., and J. Klensin, "Internationalized Domain Names Registration and Administration Guidelines for European Languages Using Cyrillic", RFC 5992, DOI 10.17487/RFC5992, October 2010, <https://www.rfc-editor.org/info/rfc5992>.

[RFC6452] Faltstrom, P., Ed. and P. Hoffman, Ed., "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA) - Unicode 6.0", RFC 6452, DOI 10.17487/RFC6452, November 2011, <https://www.rfc-editor.org/info/rfc6452>.

[RFC6912] Sullivan, A., Thaler, D., Klensin, J., and O. Kolkman, "Principles for Unicode Code Point Inclusion in Labels in the DNS", RFC 6912, DOI 10.17487/RFC6912, April 2013, <https://www.rfc-editor.org/info/rfc6912>.

Appendix A. Change Log

RFC Editor: Please remove this appendix before publication.

A.1. Changes from version -00 (2017-03-11) to -01

- Added Acknowledgments and adjusted references.
- Filled in Section 4 with updates to respond to errata.
- Added Section 5 to discuss relationships to other documents.
- Modified the Abstract to note specifically updated documents.
- Several small editorial changes and corrections.

Authors' Addresses

John C Klensin
1770 Massachusetts Ave, Ste 322
Cambridge, MA 02140
USA

Phone: +1 617 245 1457
Email: john-ietf@jck.com

Asmus Freytag
ASMUS, Inc.

Email: asmus@unicode.org