Proposal for a Latin Script Root Zone LGR

LGR Version: 5
Date: 23 September 2021
Document version: 7.1
Authors: Latin GP

Table of Contents

1. General Information 3
2. Script for which the LGR is proposed 3
3. Background on Script and Principal Languages Using It 5
3.1. Principal Languages Using Latin Script 5
3.2. Geographic Territories or Countries with Significant Latin Script User Communities 6
3.3. Related Scripts 6
4. Overall Development Process and Methodology 7
5. Repertoire 8
5.1. Definitions 8
5.2. Principles for Developing Repertoire 9
5.2.1. Inclusion Principles 9
5.2.2. Exclusion Principles 9
5.3. Included code points 10
5.3.1. Note on Combining Marks 31
5.3.2. Note on Caron with Letters d, l, and t 32
5.4. Excluded Code Points 32
5.4.1. Other Excluded Letters 34
6. Variants 36
6.1. Principles for In-Script Variants 36
6.1.1. Distinguishing Visual from Non-Visual Variants 37
6.1.2. Visual Variants 38
6.1.3. Non-Visual Variant: Shape of Base Characters 39
6.1.4. Non-Visual Variant: Spacing of Base Characters 40

6.1.5. Non-Visual Variant: IDNA 2003 Compatibility 40
6.1.6. Non-Visual Variant: Shape of Diacritics 41
6.1.7. Non-Visual Variant: Stacking of Diacritics 41
6.2. Methodology for Developing Variants 42
6.2.1. In-Script Variants 42
6.2.2. Cross-Script Variants 44
6.3. Variant Sets 45
6.3.1. In-Script Variants 45
6.3.1.1. Variant Pairs with Diacritics: Breve and Caron 45
6.3.1.2. Variant Pairs with Diacritics: Tilde and Macron 46
6.3.1.3. Variant Pairs with Diacritics: Grave and Hook Above 47
6.3.1.4. Variant Pairs with Diacritics: Acute and Dot Above 47
6.3.1.5. Variant Pairs with Diacritics: Acute and Hook Above 48
6.3.1.6. Additional In-script Variant Pairs 48
6.3.2. Cross-Script Variants 50
6.3.2.1. Armenian Script 50
6.3.2.2. Cyrillic Script 51
6.3.2.3. Greek Script 55
6.3.2.4. Generic Glyphs 57
6.4. Other Considerations for Variant Analysis 58
6.4.1. URL Underlining 59
6.4.2. IDNA 2003 Compatibility 62
6.4.2.1. Latin Small Letter Sharp S 62
6.4.2.2. Latin Small Letter Dotless I 63
In-Script Variant Mapping Types 64
6.5. Variant Due to Transitivity 64
6.6. Additional Discussion on Variants 65
6.7. Complete Variant Sets 65
7. Whole Label Evaluation Rules (WLE) and contextual Rules 77
8. Contributors 77
9. References 78

1. General Information

The purpose of this document is to give an overview of the proposed LGR in the XML format and the rationale behind the decisions taken. It includes a discussion of relevant features of the script, the communities or languages using it, the process and methodology used, and information on the contributors. The formal specification of the LGR can be found in the accompanying XML document:

proposal-latin-lgr-23sep21-en.xml

Labels for testing can be found in the accompanying text document:

latin-test-labels-23sep21-en.txt

All the appendices to the document can be found in the accompanying documents:

Appendix A - Updated MSR during Latin GP work.pdf
Appendix B - Table of Languages Used to Develop Latin Script Repertoire.pdf
Appendix C - Repertoire Table Grouped by Glyph.pdf
Appendix D - Variants Analysis.pdf
Appendix D.1 - Shape of Base Characters.pdf
Appendix D.2 - Spacing of Base Characters.pdf
Appendix D.3 - Shape of Diacritics.pdf
Appendix D.4 - Stacking of Diacritics.pdf
Appendix D.5 - IDNA 2003 Compatibility.pdf
Appendix D.6 - Underlining Evaluation Process.pdf
Appendix D.7 - Generic Glyphs.pdf
Appendix D.8 - Caron Above.pdf
Appendix D.9 - Cross-script Variants.pdf
Appendix E - Visually Confusable Glyphs.pdf

2. Script for which the LGR is proposed

The Latin script has the following specifications:

- ∉ ISO 15924 code: Latn
- ∉ ISO 15924 no.: 215
- ∉ ISO 15924 English Name: Latin

Native name of the script:

- It is written differently in different languages. A partial list of script names in different languages is given below:

  Latin (English, French)
  Latino (Italian, Portuguese)

Latín (Spanish)
Latinica (Croatian, Serbian)
Kịch bản latin (Vietnamese)
Umbhalo we-latin (Zulu)

Maximal Starting Repertoire (MSR) version: MSR-5

As per the Procedure to Develop and Maintain the Label Generation Rules for the DNS Root Zone in Respect of IDNA Labels (referred to simply as [Procedure] in the following), only code points included in the latest version of the Maximal Starting Repertoire (currently version 5 and referred to simply as [MSR] in the following) were considered.

The set of code points in the Latin script, as specified by [MSR], contains 347 selected code points, i.e., 327 letters and 20 Combining Diacritical Marks. Code points are from the following Unicode ranges as listed in table 1 below. [MSR] excludes the Unicode ranges listed in table 2 below.

Table 1. Unicode ranges included in [MSR].

Latin Script	Range of Unicode code points
Controls and Basic Latin	U+0061 – U+007A
Controls and Latin-1 Supplement	U+00DF - U+00F6 U+00F8 - U+00FF
Latin Extended-A	U+0101 – U+017F
Latin Extended-B	U+0180 – U+024F
IPA Extensions	U+0250 – U+02AF
Combining Diacritical Marks	U+0300 – U+036F
Combining Diacritical Marks Supplement	U+1DC0 – U+1DFF
Latin Extended Additional	U+1E00 – U+1EFF
Latin Extended-C	U+2C60 – U+2C7F
Latin Extended-D	U+A7B9

Table 2. Unicode ranges excluded from [MSR]

Latin Script	Range of Unicode code points
Latin Extended-D; technical use (phonetic)/obsolete/punctuation	U+A720 – U+A7FF
Latin Ligatures; compatibility characters not PVALID in IDNA 2008	U+FB00 – U+FB0F
Full-width Latin Letters; compatibility characters not PVALID in IDNA 2008	U+FF00 – U+FF5E

When a single, precomposed code point is equivalent to the combination of letter code point and a diacritic mark code point, only the precomposed code point may be used, per [IDNA 2008]. Furthermore, only lower-case letters are considered in creating the repertoire, as upper-case ones may not be used in IDNs, per [IDNA 2008]. [IDNA 2008] replaces the older IDNA version, [IDNA 2003].

3. Background on Script and Principal Languages Using It

The Latin script[1] is a major writing system of the world today. It is the most widely used in terms of number of languages and number of speakers, with circa 70% of the world's readers and writers making use of this script[2] [Wikipedia-Latin script].

3.1. Principal Languages Using Latin Script

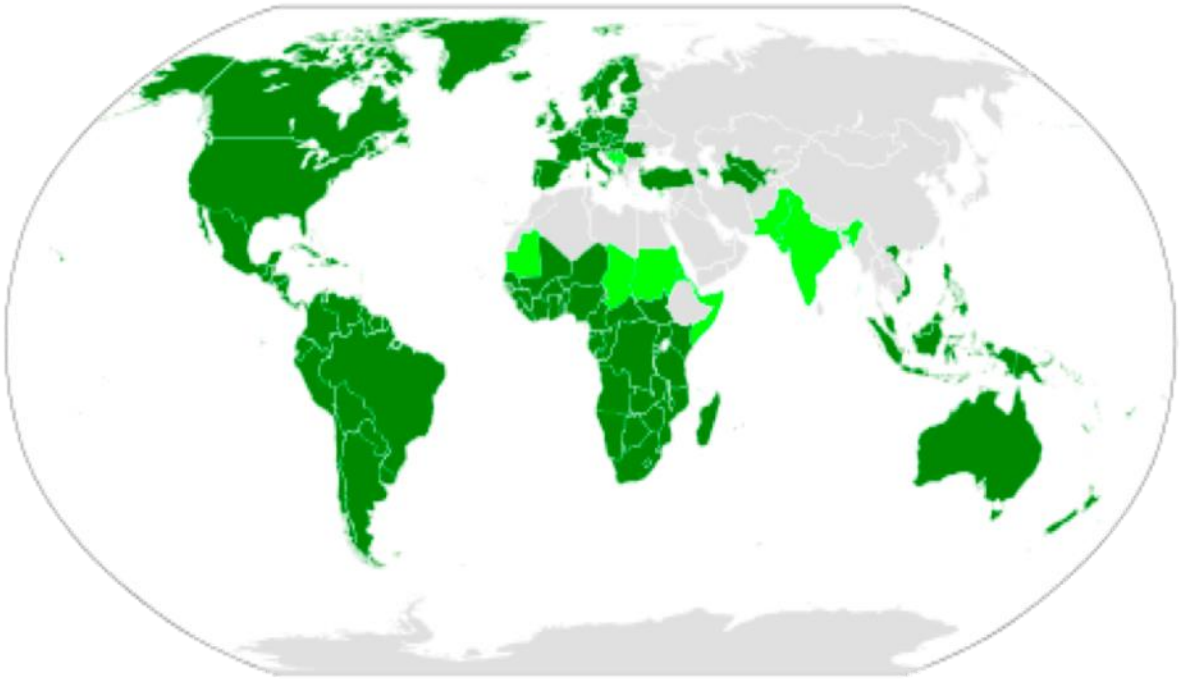
The list of languages taken into consideration contains relevant data for 455 languages using Latin script. The table with languages using Latin script was derived using data from http://www.omniglot.com/writing/langalph.htm and https://www.ethnologue.com/browse/names.

Table with 212 considered languages is in Appendix B of this document. See Section 4 for details.

[1] *Script* is used here to indicate the whole writing system including basic letters, ligatures and diacritics. See also RFC 6365 and ISO 15924.
[2] However, several orthographies on the basis of different scripts are frequently used simultaneously, both historically and contemporarily.

3.2. Geographic Territories or Countries with Significant Latin Script User Communities

Per Wikipedia the distribution of the Latin script on the world map is:



Dark green marks countries where the Latin script is the sole main script.
Light green marks countries where Latin co-exists with other scripts.
Grey marks areas, in which the Latin script is not used or used only unofficially for a second language.

3.3. Related Scripts

Latin GP observes that the following scripts are related:

1. Cyrillic
2. Greek
3. Armenian

Latin, Cyrillic and Armenian are all derived from Greek.

4. Overall Development Process and Methodology

The work has been done according to the work plan given in "Proposal for the Generation Panel (GP) for the Latin Script Label Generation Ruleset (LGR) for the Root Zone".
The panel formed two working groups:

- Repertoire WG
- Variant WG

which worked in parallel.

The first task for each group was to define the Principles for developing Repertoire and the Principles for developing Variants. The principles were sent to the Integration Panel for comments and suggestions and were also offered for public unofficial comment. Comments from the Integration Panel were encompassed in final version of the Principles.

During the Repertoire definition phase, the Latin Generation Panel reviewed and processed 181 languages with EGIDS level 1 through 4, and 319 languages with EGIDS Level 5 which have more than 1,000,000 speakers. The processed languages are listed in Appendix B.

The Latin Generation Panel used [MSR] as the starting point and after processing 212 languages the Latin GP found:

1. 197 code points verified,
2. 21 code point sequences (defined below) detected,
3. 1 code point sequence of Latin MSR code points, "ss" (that is Latin Small Letter S twice in sequence) has been added to the repertoire for technical reasons.

The panel also found that some languages use letters matching code points outside [MSR]. In some cases, these code points were rejected. In 6 cases, the panel made successful requests for inclusion of additional Code Points in [MSR]. This is described in more detail in Appendix A.

The second phase of Latin GP work was mainly devoted to defining in-script and cross-script Variants.

5. Repertoire

Based on the discussions within the GP, the principles for inclusion and exclusion of code points in the Repertoire are as follows.

5.1. Definitions

Language: The present document and its principles deal with any language making use of Latin script[3] today. Languages are restricted to natural human languages in active use. Both the socio-political situation (such as the political or legal status of a language in a country or community) and the socio-linguistic roles of languages in society (such as the absolute or relative frequency of use) are explicitly not considered for the current purposes. Super- or sub-units of languages, such as dialect, regiolect (a dialect spoken in a particular geographical region), or language clusters, are all considered equivalent to language. However, notions such as official language, national language, standard language and vernacular, are not considered at all in determining whether something is a language.

Letter Code Point is a Unicode code point with General Category property value of Lx (Lu, Ll, Lt, Lm, Lo), as defined in the Unicode Character Database.

Mark Code Point is a Unicode code point with General Category property value of Mx (Mn, Mc, Me), as defined in the Unicode Character Database.

Code Point Sequence is a sequence of two or more Code Points (that is, a glyph formed by a Letter Code Point followed by one or more combining diacritic Mark Code Point(s). This is used for cases when Unicode does not include a single Code Point for the glyph).

Established contemporary use of a letter means it is in active use by a community today. Such use may be demonstrated by, for example, educational resources, published material, media, or other materials and sources. This does not depend on their material or non-material form, such as handwritten or typed manuscripts or digitally produced text. There may be multiple sources for acquiring such evidence, including (but not limited to) the following:

1. Members of Language communities,
2. Members of the Latin GP,
3. Other experts

[3] The Latin script is also known as Roman script in academic literature.

4. Language tables submitted by ccTLD in the context of IDNA 2008 in the IANA repository, and
5. Published standards (e.g., by a language authority or any other national or international body).

5.2. Principles for Developing Repertoire

5.2.1. Inclusion Principles

Based on the MSR-5, if a Code Point is included as part of a label, the Code Point cannot be retracted in future revisions of the LGR. All applicable criteria must be met to include a Code Point.

- Only languages which have a rating of levels of 0-4 under the Expanded Graded Intergenerational Disruption Scale (EGIDS) are considered as supporting the inclusion of a Code Point. Languages with EGIDS 5 may be included in special cases where there is additional evidence that it is in widespread use, notwithstanding its formal EGIDS rating. For these, a threshold of 1 million native speakers was used.
- Code Points may only be included if they have established contemporary use in one or more of the languages considered.
- If the Code Point in question is a Mark Code Point, then it can only be included in its context. That is, a Mark Code Point is included as part of a sequence consisting of a Lower Letter (Ll) or Other Letter (Lo) and the subsequent mark or marks. (See Section 5.3.1)
- Any combination of Code Points is defined by its sequence. To be included, a sequence must be supported by some included language in the same way as a separate Code Point of type Ll or Lo.
- For Latin script, where a precomposed alternative exists, it is used (in other words, NFC[4] Form is always used).

5.2.2. Exclusion Principles

A Code Point is excluded if at least one of these exclusion principles is met. (If a Code Point can neither be included nor excluded on the basis of these principles, the Code Point is automatically excluded from the proposed LGR for Latin Script, per RFC 6912.)

1. The Code Point is DISALLOWED or UNASSIGNED by IDNA 2008 protocol.
2. The Code Point presents a security or stability issue which cannot be resolved at any other stage of the analysis (e.g., stage of determining Code Points, variants, Contextual Rules or Whole Label Evaluation Rules).

[4] See https://unicode.org/reports/tr15/

3. The Code Point is either deprecated or not recommended for use in Unicode Standard -- unless it meets all of the applicable inclusion criteria, with no alternative Code Point or Code Point sequence.
4. The Code Point is used exclusively in a subset of textual genres, such as technical or religious texts, and is not otherwise used as described in Section 2 above.
5. The Code Point is predominantly used in one of the following functions, apart from any other uses in orthography:
    a. Formatting character or mark
    b. Numerical digit
    c. Punctuation mark
    d. Honorific mark or symbol
    e. Mathematical symbol
6. The Code Point is difficult to distinguish from a Code Point which fits the criteria in #5. See Section 5.4.

5.3. Included code points

The table below lists the code points and sequences proposed for inclusion in the root zone LGR for the Latin script. The table excludes sequence “ss” which is needed for variant definitions but redundant from a repertoire perspective.

The table also lists examples of languages using the code point and their EGIDS rating. All references for specific code points found during the review of languages contributing to the repertoire are included. This table is sorted by Unicode column. (A table with the same data, sorted by glyph, can be found in Appendix C.) The list of references supporting inclusion of code points is in Section 9.

Table 3. Code Points Included in the Repertoire of Latin Script LGR.

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
1	0061	a	LATIN SMALL LETTER A	Basic Latin	[99]
2	0061 + 0331	a̲	LATIN SMALL LETTER A + COMBINING MACRON BELOW	Nuer (4)	[146], [129]
3	0062	b	LATIN SMALL LETTER B	Basic Latin	[99]
4	0063	c	LATIN SMALL LETTER C	Basic Latin	[99]
5	0064	d	LATIN SMALL LETTER D	Basic Latin	[99]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
6	0065	e	LATIN SMALL LETTER E	Basic Latin	[99]
7	0065 + 0331	e̱	LATIN SMALL LETTER E + COMBINING MACRON BELOW	Nuer (4)	[146]
8	0066	f	LATIN SMALL LETTER F	Basic Latin	[99]
9	0067	g	LATIN SMALL LETTER G	Basic Latin	[99]
10	0067 + 0303	g̃	LATIN SMALL LETTER G + COMBINING TILDE	Guarani (1)	[142], [143]
11	0068	h	LATIN SMALL LETTER H	Basic Latin	[99]
12	0069	i	LATIN SMALL LETTER I	Basic Latin	[99]
13	0069 + 0331	i̱	LATIN SMALL LETTER I + COMBINING MACRON BELOW	Nuer (4)	[146]
14	006A	j	LATIN SMALL LETTER J	Basic Latin	[99]
15	006B	k	LATIN SMALL LETTER K	Basic Latin	[99]
16	006C	l	LATIN SMALL LETTER L	Basic Latin	[99]
17	006D	m	LATIN SMALL LETTER M	Basic Latin	[99]
18	006D + 0327	m̧	LATIN SMALL LETTER M + COMBINING CEDILLA	Marshallese (1)	[213], [136], [214]
19	006E	n	LATIN SMALL LETTER N	Basic Latin	[99]
20	006E + 0304	n̄	LATIN SMALL LETTER N + COMBINING MACRON	Raga (Hano) (3) Marshallese (1)	[200], [213], [136]
21	006E + 0308	n̈	LATIN SMALL LETTER N + COMBINING DIAERESIS	Malagasy (1)	[276]
22	006F	o	LATIN SMALL LETTER O	Basic Latin	[99]
23	006F + 0327	o̧	LATIN SMALL LETTER O + COMBINING CEDILLA	Marshallese (1)	[136]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
24	006F + 0331	o̱	LATIN SMALL LETTER O + COMBINING MACRON BELOW	Nuer (4)	[146], [129]
25	0070	p	LATIN SMALL LETTER P	Basic Latin	[99]
26	0071	q	LATIN SMALL LETTER Q	Basic Latin	[99]
27	0072	r	LATIN SMALL LETTER R	Basic Latin	[99]
28	0072 + 0303	r̃	LATIN SMALL LETTER R WITH COMBINING TILDE	Hausa (2)	[147]
29	0073	s	LATIN SMALL LETTER S	Basic Latin	[99]
30	0074	t	LATIN SMALL LETTER T	Basic Latin	[99]
31	0075	u	LATIN SMALL LETTER U	Basic Latin	[99]
32	0076	v	LATIN SMALL LETTER V	Basic Latin	[99]
33	0077	w	LATIN SMALL LETTER W	Basic Latin	[99]
34	0078	x	LATIN SMALL LETTER X	Basic Latin	[99]
35	0079	y	LATIN SMALL LETTER Y	Basic Latin	[99]
36	007A	z	LATIN SMALL LETTER Z	Basic Latin	[99]
37	00DF	ß	LATIN SMALL LETTER SHARP S	German (1)	[119]
38	00E0	à	LATIN SMALL LETTER A WITH GRAVE	Italian (1) French (1) Galician (2) Wolof (4)	[114]. [130], [131], [106], [132]
39	00E1	á	LATIN SMALL LETTER A WITH ACUTE	Spanish (1) Czech (1) Icelandic (1) Faroese (2) Chuukese (2) Galician (2) Lule Sámi (2) Northern Sámi (2)	[100], [101], [102], [103], [105], [106], [107], [108]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
40	00E2	â	LATIN SMALL LETTER A WITH CIRCUMFLEX	Vietnamese (1) Romanian (1) Skolt Sami (2) French (1) Galician (2) West Frisian (2) Friulian (4) Xavante (4)	[109], [110], [113], [114], [106], [115], [116], [117], [275]
41	00E3	ã	LATIN SMALL LETTER A WITH TILDE	Umbundu (3) Guarani (1) Nauruan (3) Khoekhoe (4)	[141], [142], [143], [144], [145]
42	00E4	ä	LATIN SMALL LETTER A WITH DIAERESIS	German (1) Finnish (1) Turkmen (1) Estonian (1) Swedish (1) Lule Sámi (2) Yapese (2) Dinka (4) Kaqchikel (4) Bashkir (4) Alsatian (5) Nuer (4)	[119], [120], [121], [122], [123], [107], [124], [125], [126], [127], [128], [129]
43	00E5	å	LATIN SMALL LETTER A WITH RING ABOVE	Danish (1) Finnish (1) Chamorro (1) Swedish (1) Lule Sámi (2)	[139], [120], [140], [123], [107]
44	00E6	æ	LATIN SMALL LETTER AE	Danish (1) Icelandic (1) Faroese (2)	[139], [102], [103]
45	00E7	ç	LATIN SMALL LETTER C WITH CEDILLA	Turkish (1) Turkmen (1) Kurdish (2) French (1) Azerbaijani (1) Basque (1) Galician (2)	[157], [121], [158], [114], [159], [160], [161], [106], [116], [127]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
				Friulian (4) Bashkir (4)	
46	00E8	è	LATIN SMALL LETTER E WITH GRAVE	French (1) Italian (1) Afrikaans (1) Haitian Creole (1) French (1)	[114], [130], [175], [182], [183]
47	00E9	é	LATIN SMALL LETTER E WITH ACUTE	French (1) Italian (1) Spanish (1) Czech (1) Icelandic (1) Chuukese (2) Galician (2) Wolof (4) Xavante (4) West Frisian (2)	[114], [130], [100], [101], [102], [105], [106], [132], [117], [275], [115]
48	00EA	ê	LATIN SMALL LETTER E WITH CIRCUMFLEX	French (1) Tswana (1) Afrikaans (1) Vietnamese (1) Kurdish (2) West Frisian (2) Friulian (4)	[114], [173], [174], [175], [109], [158], [115], [116]
49	00EB	ë	LATIN SMALL LETTER E WITH DIAERESIS	Afrikaans (1) Albanian (1) French (1) Uyghur (2) Yapese (2) Wolof (4) Drehu (4) Kaqchikel (4) West Frisian (2) Nuer (4)	[175], [176], [177], [114], [105], [179], [124], [132], [180], [126], [115], [129]
50	00EC	ì	LATIN SMALL LETTER I WITH GRAVE	Italian (1)	[130], [206], [208]
51	00ED	í	LATIN SMALL LETTER I WITH ACUTE	Spanish (1) Czech (1)	[100], [101], [102], [103], [106], [127]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
				Icelandic (1) Faroese (2) Galician (2) Bashkir (4)	
52	00EE	î	LATIN SMALL LETTER I WITH CIRCUMFLEX	Afrikaans (1) Romanian (1) Kurdish (2) French (1) Friulian (4)	[175], [110], [158], [114], [116]
53	00EF	ï	LATIN SMALL LETTER I WITH DIAERESIS	Afrikaans (1) French (1) Kaqchikel (4) Dinka (4) West Frisian (2)	[175], [114], [126], [125], [115]
54	00F0	ð	LATIN SMALL LETTER ETH	Faroese (2) Icelandic (1)	[103], [102]
55	00F1	ñ	LATIN SMALL LETTER N WITH TILDE	Spanish (1) Fula (3) Chamorro (1) Filipino (1) Guarani (1) Chavacano (4) Basque (1) Galician (2) Iloco (3) Quechua (3) Cape Verdean Creole (4) Waray-Waray (3) Wolof (4) Nauruan (3) Lozi (4) Bashkir (4) Marshallese (1) Mandinka (5) Igbo (2)	[221], [149],[222], [142], [143], [223], [160], [106], [224], [225], [226], [227], [228], [132], [144], [229], [127], [136], [197], [205]
56	00F2	ò	LATIN SMALL LETTER O WITH GRAVE	Italian (1) Haitian Creole (1)	[130], [182], [183]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
57	00F3	ó	LATIN SMALL LETTER O WITH ACUTE	Spanish (1) Polish (1) Czech (1) Icelandic (1) Chuukese (2) Galician (2) Wolof (4)	[100], [152], [101], [102], [105], [106], [132]
58	00F4	ô	LATIN SMALL LETTER O WITH CIRCUMFLEX	Tswana (1) Afrikaans (1) Vietnamese (1) French (1) Northern Sotho (1) West Frisian (2) Galician (2) Friulian (4) Xavante (4)	[173], [174], [175], [109], [114], [230], [115], [106], [116], [117], [275]
59	00F5	õ	LATIN SMALL LETTER O WITH TILDE	Estonian (1) Skolt Sami (2) Umbundu (3) Guarani (1) Nauruan (3) Xavante (4) Khoekhoe (4)	[122], [113], [141], [142], [143], [144], [117], [275], [145]
60	00F6	ö	LATIN SMALL LETTER O WITH DIAERESIS	German (1) Finnish (1) Afrikaans (1) Turkish (1) Swedish (1) Uygur (2) Yapese (2) Drehu (4) Kaqchikel (4) Dinka (4) Bashkir (4) Chechen (2) 1992 Version West Frisian (2) Nuer (4)	[119], [120], [175], [157], [123], [179], [124], [180], [126], [125], [127], [232], [115], [129]
61	00F8	ø	LATIN SMALL LETTER O WITH STROKE	Danish (1) Faroese (2)	[139], [103]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
62	00F9	ù	LATIN SMALL LETTER U WITH GRAVE	Italian (1) French (1) Papiamento (1)	[130], [114],[206], [245], [246], [253]
63	00FA	ú	LATIN SMALL LETTER U WITH ACUTE	Spanish (1) Czech (1) Icelandic (1) Faroese (2) Chuukese (2) West Frisian (2) Galician (2)	[100], [101], [102], [103], [105], [115], [106]
64	00FB	û	LATIN SMALL LETTER U WITH CIRCUMFLEX	Afrikaans (1) Kurdish (2) French (1) Miskito (2) West Frisian (2) Friulian (4) Zazaki (4)	[175], [158], [114], [243], [115], [116], [202]
65	00FC	ü	LATIN SMALL LETTER U WITH DIAERESIS	German (1) Spanish (1) Afrikaans (1) Turkish (1) Swedish (1) French (1) Azeri (1) Basque (1) Galician (2) Uygur (2) Kaqchikel (4) Bashkir (4)	[119], [100], [175], [157], [123], [114], [159], [161], [106], [179], [126], [127]
66	00FD	ý	LATIN SMALL LETTER Y WITH ACUTE	Turkmen (1) Czech (1) Icelandic (1) Faroese (2) Guarani (1)	[121], [101], [102], [103], [142], [143]
67	00FE	þ	LATIN SMALL LETTER THORN	Icelandic (1)	[102]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
68	00FF	ÿ	LATIN SMALL LETTER Y WITH DIAERESIS	French (1)	[114], [253], [257]
69	0101	ā	LATIN SMALL LETTER A WITH MACRON	Latvian (1) Tongan (1) Hawaiian (2) Marshallese (1)	[133], [134], [135], [136]
70	0103	ă	LATIN SMALL LETTER A WITH BREVE	Vietnamese (1) Romanian (1)	[109], [110]
71	0105	ą	LATIN SMALL LETTER A WITH OGONEK	Polish (1) Lithuanian (1)	[137], [138]
72	0107	ć	LATIN SMALL LETTER C WITH ACUTE	Croatian (1) Serbian (1) Polish (1)	[150], [151], [152]
73	0109	ĉ	LATIN SMALL LETTER C WITH CIRCUMFLEX	Esperanto (3)	[255]
74	010B	ċ	LATIN SMALL LETTER C WITH DOT ABOVE	Maltese (1)	[163]
75	010D	č	LATIN SMALL LETTER C WITH CARON	Croatian (1) Serbian (1) Latvian (1) Slovak (1) Northern Sámi (2) Lithuanian (1)	[150], [151], [133], [153], [108], [154]
76	010F	ď	LATIN SMALL LETTER D WITH CARON	Czech (1) Slovak (1)	[101], [153]
77	0111	đ	LATIN SMALL LETTER D WITH STROKE	Croatian (1) Serbian (1) Vietnamese (1) Northern Sámi (2) Brahui (5)	[150], [151], [109], [108], [168]
78	0113	ē	LATIN SMALL LETTER E WITH MACRON	Latvian (1) Hawaiian (2)	[133], [135], [134], [184]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
				Tongan (1) Minangkabau (5)	
79	0117	ė	LATIN SMALL LETTER E WITH DOT ABOVE	Lithuanian (1)	[138], [154]
80	0119	ę	LATIN SMALL LETTER E WITH OGONEK	Polish (1) Palauan (2) Lithuanian (1)	[152], [185], [138], [154]
81	011B	ě	LATIN SMALL LETTER E WITH CARON	Czech (1) Sorbian (4)	[101], [172]
82	011D	ĝ	LATIN SMALL LETTER G WITH CIRCUMFLEX	Esperanto (3)	[255]
83	011F	ğ	LATIN SMALL LETTER G WITH BREVE	Turkish (1) Tatar (2) Azeri (1) Bashkir (4) Zaza (5)	[157], [201], [159], [127], [202]
84	0121	ġ	LATIN SMALL LETTER G WITH DOT ABOVE	Maltese (1)	[163]
85	0123	ģ	LATIN SMALL LETTER G WITH CEDILLA	Latvian (1) Brahui (5)	[133], [168]
86	0125	ĥ	LATIN SMALL LETTER H WITH CIRCUMFLEX	Esperanto (3)	[255]
87	0127	ħ	LATIN SMALL LETTER H WITH STROKE	Maltese (1)	[163]
88	0129	ĩ	LATIN SMALL LETTER I WITH TILDE	Guarani (1) Cubeo (3) Khoekhoe (4) Kikuyu (5)	[142], [143], [186], [145], [209]
89	012B	ī	LATIN SMALL LETTER I WITH MACRON	Latvian (1) Lithuanian (1) Hawaiian (2) Tongan (1)	[133], [138], [135], [134]
90	012F	į	LATIN SMALL LETTER I WITH OGONEK	Lithuanian (1)	[154]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
91	0131	ı	LATIN SMALL LETTER DOTLESS I	Turkish (1) Tatar (2) Azeri (1)	[157], [203], [201], [159]
92	0135	ĵ	LATIN SMALL LETTER J WITH CIRCUMFLEX	Esperanto (3)	[255]
93	0137	ķ	LATIN SMALL LETTER K WITH CEDILLA	Latvian (1)	[133]
94	013A	ĺ	LATIN SMALL LETTER L WITH ACUTE	Slovak (1)	[153]
95	013C	ļ	LATIN SMALL LETTER L WITH CEDILLA	Latvian (1) Marshallese (1) Brahui (5)	[133], [213], [214], [168]
96	013E	ľ	LATIN SMALL LETTER L WITH CARON	Slovak (1)	[153]
97	0142	ł	LATIN SMALL LETTER L WITH STROKE	Polish (1)	[152]
98	0144	ń	LATIN SMALL LETTER N WITH ACUTE	Polish (1) Lule Sámi (2) Sorbian (4) Brahui (5)	[152], [107], [172], [168]
99	0146	ņ	LATIN SMALL LETTER N WITH CEDILLA	Latvian (1) Marshallese (1)	[133], [136]
100	0148	ň	LATIN SMALL LETTER N WITH CARON	Turkmen (1) Czech (1) Slovak (1)	[121], [101], [153]
101	014B	ŋ	LATIN SMALL LETTER ENG	Inari Sami (2) Dagaare Burkina Faso (4) Dagbani (Dagomba) (4) Northern Sami (2) Ewondo (3) Luganda (3) Wolof (4) Adzera (4)	[188], [148], [189], [108], [190], [191], [132], [192], [146], [193], [125], [194], [170], [195], [196], [197], [198], [199], [129]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
				Nuer (4) Ga (4) Dinka (4) Duala (3) Ewe (3) Soga (5) Alur (5) Mandinka (5) Acholi (5) Bambara (4)	
102	014D	ō	LATIN SMALL LETTER O WITH MACRON	Hawaiian (2) Marshallese (1) Tongan (1)	[135], [136], [134]
103	0151	ő	LATIN SMALL LETTER O WITH DOUBLE ACUTE	Hungarian (1)	[233], [234]
104	0153	œ	LATIN SMALL LIGATURE OE	French (1)	[114], [253]
105	0155	ŕ	LATIN SMALL LETTER R WITH ACUTE	Slovak (1) Brahui (5)	[153], [168]
106	0159	ř	LATIN SMALL LETTER R WITH CARON	Czech (1) Sorbian (4)	[101], [172]
107	015B	ś	LATIN SMALL LETTER S WITH ACUTE	Polish (1) Montenegrin (1)	[152], [258]
108	015D	ŝ	LATIN SMALL LETTER S WITH CIRCUMFLEX	Esperanto (3)	[255]
109	015F	ş	LATIN SMALL LETTER S WITH CEDILLA	Turkish (1) Turkmen (1) Kurdish (2) Tatar (2) Azeri (1) Bashkir (4) Brahui (5) Zaza (5)	[157], [121], [158], [201], [159], [127], [168], [202]
110	0161	š	LATIN SMALL LETTER S WITH CARON	Tswana (1) Croatian (1) Serbian (1) Latvian (1)	[174], [150], [151], [133], [230], [108], [154]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
				Northern Sotho (1) Northern Sami (2) Lithuanian (1)	
111	0165	ť	LATIN SMALL LETTER T WITH CARON	Czech (1) Slovak (1)	[101], [153]
112	0167	ŧ	LATIN SMALL LETTER T WITH STROKE	Northern Sami (2) Brahui (5)	[108], [168]
113	0169	ũ	LATIN SMALL LETTER U WITH TILDE	Umbundu (3) Guarani (1) Nauruan (3) Khoekhoe (4) Kikuyu (5)	[141], [142], [143], [144], [145], [209]
114	016B	ū	LATIN SMALL LETTER U WITH MACRON	Latvian (1) Hawaiian (2) Lithuanian (1) Marshallese (1) Tongan (1)	[133], [135], [138], [154], [136], [134]
115	016D	ŭ	LATIN SMALL LETTER U WITH BREVE	Esperanto (3)	[255]
116	016F	ů	LATIN SMALL LETTER U WITH RING ABOVE	Czech (1)	[101]
117	0171	ű	LATIN SMALL LETTER U WITH DOUBLE ACUTE	Hungarian (1)	[233], [234]
118	0173	ų	LATIN SMALL LETTER U WITH OGONEK	Lithuanian (1)	[154], [138]
119	0175	ŵ	LATIN SMALL LETTER W WITH CIRCUMFLEX	Chichewa (3) Welsh (2)	[247], [256]
120	0177	ŷ	LATIN SMALL LETTER Y WITH CIRCUMFLEX	Welsh (2)	[256]
121	017A	ź	LATIN SMALL LETTER Z WITH ACUTE	Polish (1) Brahui (5) Sorbian (4) Montenegrin (1)	[152], [252], [168], [172], [258]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
122	017C	ż	LATIN SMALL LETTER Z WITH DOT ABOVE	Polish (1) Maltese (1)	[152], [163]
123	017E	ž	LATIN SMALL LETTER Z WITH CARON	Lithuanian (1) Croatian (1) Serbian (1) Turkmen (1) Latvian (1) Slovak (1) Northern Sami (2) Chechen (2) 1925 Version	[154], [150], [151], [121], [133], [153], [108], [232]
124	0188	ƈ	LATIN SMALL LETTER C WITH HOOK	Serer (5)	[277]
125	0192	ƒ	LATIN SMALL LETTER F WITH HOOK	Ewe (3)	[170]
126	0199	ƙ	LATIN SMALL LETTER K WITH HOOK	Hausa (2)	[147]
127	01A1	ơ	LATIN SMALL LETTER O WITH HORN	Vietnamese (1)	[109]
128	01A5	ƥ	LATIN SMALL LETTER P WITH HOOK	Serer (5)	[277]
129	01AD	ƭ	LATIN SMALL LETTER T WITH HOOK	Serer (5)	[277]
130	01B0	ư	LATIN SMALL LETTER U WITH HORN	Vietnamese (1)	[109]
131	01B4	ƴ	LATIN SMALL LETTER Y WITH HOOK	Dagaare-Burkina Faso (4) Fula (3)	[148], [251], [149]
132	01DD	ǝ	LATIN SMALL LETTER TURNED E	Kanuri (3)	[240]
133	01E7	ǧ	LATIN SMALL LETTER G WITH CARON	Skolt Sami (2)	[113]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
134	01E9	ǩ	LATIN SMALL LETTER K WITH CARON	Skolt Sami (2)	[113]
135	01EF	ǯ	LATIN SMALL LETTER EZH WITH CARON	Skolt Sami (2)	[113]
136	0219	ș	LATIN SMALL LETTER S WITH COMMA BELOW	Romanian (1)	[110]
137	021B	ț	LATIN SMALL LETTER T WITH COMMA BELOW	Romanian (1)	[110]
138	024D	ɍ	LATIN SMALL LETTER R WITH STROKE	Kanuri (3)	[240]
139	0253	ɓ	LATIN SMALL LETTER B WITH HOOK	Hausa (2) Dagaare-Burkina Faso (4) Fula (3)	[147], [148], [250]
140	0254	ɔ	LATIN SMALL LETTER OPEN O	Dagaare - Burkina Faso (4) Dagbani (Dagomba) (4) Lingala (2) Akan (3) Ewondo (3) Fon (3) Nuer (4) Ga (4) Duala (3) Ewe (3) Nuer (4)	[148], [189], [236], [237], [190], [169], [146], [193], [194], [170], [129]
141	0254 + 0308	ɔ̈	LATIN SMALL LETTER OPEN O + COMBINING DIAERESIS	Dinka (4)	[125]
142	0254 + 0331	ɔ̱	LATIN SMALL LETTER OPEN O + COMBINING MACRON BELOW	Nuer (4)	[129], [146]
143	0256	ɖ	LATIN SMALL LETTER D WITH TAIL	Fon (3) Ewe (3)	[169], [170]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
144	0257	ɗ	LATIN SMALL LETTER D WITH HOOK	Hausa (2) Fula (3)	[147], [149], [250]
145	0259	ə	LATIN SMALL LETTER SCHWA	Azeri, Azerbaijani (1) Ewondo (3) Ewe (3) Bugis (3)	[159], [190], [170], [241]
146	025B	ɛ	LATIN SMALL LETTER OPEN E	Dagaare - Burkina Faso (4) Lingala (2) Akan (3) Ewondo (3) Dagbani (Dagomba) (4) Fon (3) Mossi (3) Ga (4) Ewe (3) Duala (3) Bambara (4) Nuer (4)	[148], [236], [237], [190], [189], [169], [212], [238], [193], [170], [194], [199], [129]
147	025B + 0308	ɛ̈	LATIN SMALL LETTER OPEN E + COMBINING DIAERESIS	Nuer (4) Dinka (4)	[129], [146], [239], [125]
148	025B + 0331	ɛ̱	LATIN SMALL LETTER OPEN E + COMBINING MACRON BELOW	Nuer (4)	[129], [146], [239]
149	025B + 0331 + 0308	ɛ̱̈	LATIN SMALL LETTER OPEN E + COMBINING MACRON BELOW + COMBINING DIAERESIS	Nuer (4)	[129], [146], [239]
150	0260	ɠ	LATIN SMALL LETTER G WITH HOOK	Kpelle (4)	[278]
151	0263	ɣ	LATIN SMALL LETTER GAMMA	Dagbani (Dagomba) (4) Nuer (4) Dinka (4) Ewe (3) Nuer (4)	[189], [146], [125], [170], [129]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
152	0268	ɨ	LATIN SMALL LETTER I WITH STROKE	Cubeo (3) Dagbani (Dagomba) (4) HIxkaryána (4) Maasai (5)	[186], [189], [210], [211]
153	0268 + 0303	ɨ̃	LATIN SMALL LETTER I WITH STROKE + COMBINING TILDE	Cubeo (3)	[186]
154	0269	ɩ	LATIN SMALL LETTER IOTA	Dagaare - Burkina Faso (4) Mossi (3)	[148], [212]
155	0272	ɲ	LATIN SMALL LETTER N WITH LEFT HOOK	Susu (4) Zarma (4) Bambara (4)	[218], [219], [199]
156	0289	ʉ	LATIN SMALL LETTER U BAR	Cubeo (3) Maasai (5)	[186], [187], [211]
157	0289 + 0303	ʉ̃	LATIN SMALL LETTER U BAR + COMBINING TILDE	Cubeo (3)	[186], [187]
158	028B	ʋ	LATIN SMALL LETTER V WITH HOOK	Dagaare - Burkina Faso (4) Mossi (3) Ewe (3)	[148], [212], [238], [170]
159	0292	ʒ	LATIN SMALL LETTER EZH	Skolt Sami (2) Dagbani (Dagomba) (4)	[113], [189]
160	1E13	ḓ	LATIN SMALL LETTER D WITH CIRCUMFLEX BELOW	Venda (1)	[164], [257]
161	1E21	ḡ	LATIN SMALL LETTER G WITH MACRON	Raga (Hano) (3)	[200]
162	1E3D	ḽ	LATIN SMALL LETTER L WITH CIRCUMFLEX BELOW	Venda (1)	[164], [257]
163	1E45	ṅ	LATIN SMALL LETTER N WITH DOT ABOVE	Venda (1)	[164], [257]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
164	1E49	ṉ	LATIN SMALL LETTER N WITH LINE BELOW	Pitjantjatjara (4)	[220]
165	1E4B	ṋ	LATIN SMALL LETTER N WITH CIRCUMFLEX BELOW	Venda (1)	[164], [257]
166	1E63	ṣ	LATIN SMALL LETTER S WITH DOT BELOW	Yoruba (2)	[254]
167	1E6D	ṭ	LATIN SMALL LETTER T WITH DOT BELOW	Mizo (4)	[242]
168	1E71	ṱ	LATIN SMALL LETTER T WITH CIRCUMFLEX	Venda (1)	[164], [257]
169	1E8D	ẍ	LATIN SMALL LETTER X WITH DIAERESIS	Mam (4)	[248], [249]
170	1EA1	ạ	LATIN SMALL LETTER A WITH DOT BELOW	Vietnamese (1)	[109]
171	1EA3	ả	LATIN SMALL LETTER A WITH HOOK ABOVE	Vietnamese (1)	[109]
172	1EA5	ấ	LATIN SMALL LETTER A WITH CIRCUMFLEX AND ACUTE	Vietnamese (1)	[109]
173	1EA7	ầ	LATIN SMALL LETTER A WITH CIRCUMFLEX AND GRAVE	Vietnamese (1)	[109]
174	1EA9	ẩ	LATIN SMALL LETTER A WITH CIRCUMFLEX AND HOOK ABOVE	Vietnamese (1)	[109]
175	1EAB	ẫ	LATIN SMALL LETTER A WITH CIRCUMFLEX AND TILDE	Vietnamese (1)	[109]
176	1EAD	ậ	LATIN SMALL LETTER A WITH CIRCUMFLEX AND DOT BELOW	Vietnamese (1)	[109]
177	1EAF	ắ	LATIN SMALL LETTER A WITH BREVE AND ACUTE	Vietnamese (1)	[109]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
178	1EB1	ằ	LATIN SMALL LETTER A WITH BREVE AND GRAVE	Vietnamese (1)	[109]
179	1EB3	ẳ	LATIN SMALL LETTER A WITH BREVE AND HOOK ABOVE	Vietnamese (1)	[109]
180	1EB5	ẵ	LATIN SMALL LETTER A WITH BREVE AND TILDE	Vietnamese (1)	[109]
181	1EB7	ặ	LATIN SMALL LETTER A WITH BREVE AND DOT BELOW	Vietnamese (1)	[109]
182	1EB9	ẹ	LATIN SMALL LETTER E WITH DOT BELOW	Yoruba (2)	[254]
183	1EB9 + 0300	ẹ̀	LATIN SMALL LETTER E WITH DOT BELOW + COMBINING GRAVE ACCENT	Yoruba (2)	[254]
184	1EB9 + 0301	ẹ́	LATIN SMALL LETTER E WITH DOT BELOW + COMBINING ACUTE ACCENT	Yoruba (2)	[254]
185	1EBB	ẻ	LATIN SMALL LETTER E WITH HOOK ABOVE	Vietnamese (1)	[109]
186	1EBD	ẽ	LATIN SMALL LETTER E WITH TILDE	Umbundu (3) Guarani (1) Cubeo (3) Xavante (4)	[141], [142], [143], [186], [187], [117],[275]
187	1EBF	ế	LATIN SMALL LETTER E WITH CIRCUMFLEX AND ACUTE	Vietnamese (1)	[109]
188	1EC1	ề	LATIN SMALL LETTER E WITH CIRCUMFLEX AND GRAVE	Vietnamese (1)	[109]
189	1EC3	ể	LATIN SMALL LETTER E WITH CIRCUMFLEX AND HOOK ABOVE	Vietnamese (1)	[109]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
190	1EC5	ễ	LATIN SMALL LETTER E WITH CIRCUMFLEX AND TILDE	Vietnamese (1)	[109]
191	1EC7	ệ	LATIN SMALL LETTER E WITH CIRCUMFLEX AND DOT BELOW	Vietnamese (1)	[109]
192	1EC9	ỉ	LATIN SMALL LETTER I WITH HOOK ABOVE	Vietnamese (1)	[109]
193	1ECB	ị	LATIN SMALL LETTER I WITH DOT BELOW	Igbo (2)	[205]
194	1ECD	ọ	LATIN SMALL LETTER O WITH DOT BELOW	Igbo (2) Yoruba (2) Marshallese (1)	[204], [205], [254], [136], [215], [216]
195	1ECD + 0300	ọ̀	LATIN SMALL LETTER O WIT DOT BELOW + COMBINING GRAVE ACCENT	Yoruba (2)	[254]
196	1ECD + 0301	ọ́	LATIN SMALL LETTER O WITH DOT BELOW + COMBINING ACUTE ACCENT	Yoruba (2)	[254]
197	1ECF	ỏ	LATIN SMALL LETTER O WITH HOOK ABOVE	Vietnamese (1)	[109]
198	1ED1	ố	LATIN SMALL LETTER O WITH CIRCUMFLEX AND ACUTE	Vietnamese (1)	[109]
199	1ED3	ồ	LATIN SMALL LETTER O WITH CIRCUMFLEX AND GRAVE	Vietnamese (1)	[109]
200	1ED5	ổ	LATIN SMALL LETTER O WITH CIRCUMFLEX AND HOOK ABOVE	Vietnamese (1)	[109]
201	1ED7	ỗ	LATIN SMALL LETTER O WITH CIRCUMFLEX AND TILDE	Vietnamese (1)	[109]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
202	1ED9	ộ	LATIN SMALL LETTER O WITH CIRCUMFLEX AND DOT BELOW	Vietnamese (1)	[109]
203	1EDB	ớ	LATIN SMALL LETTER O WITH HORN AND ACUTE	Vietnamese (1)	[109]
204	1EDD	ờ	LATIN SMALL LETTER O WITH HORN AND GRAVE	Vietnamese (1)	[109]
205	1EDF	ở	LATIN SMALL LETTER O WITH HORN AND HOOK ABOVE	Vietnamese (1)	[109]
206	1EE1	ỡ	LATIN SMALL LETTER O WITH HORN AND TILDE	Vietnamese (1)	[109]
207	1EE3	ợ	LATIN SMALL LETTER O WITH HORN AND DOT BELOW	Vietnamese (1)	[109]
208	1EE5	ụ	LATIN SMALL LETTER U WITH DOT BELOW	Vietnamese (1) Igbo (2)	[109], [204], [205]
209	1EE7	ủ	LATIN SMALL LETTER U WITH HOOK ABOVE	Vietnamese (1)	[109]
210	1EE9	ứ	LATIN SMALL LETTER U WITH HORN AND ACUTE	Vietnamese (1)	[109]
211	1EEB	ừ	LATIN SMALL LETTER U WITH HORN AND GRAVE	Vietnamese (1)	[109]
212	1EED	ử	LATIN SMALL LETTER U WITH HORN AND HOOK ABOVE	Vietnamese (1)	[109]
213	1EEF	ữ	LATIN SMALL LETTER U WITH HORN AND TILDE	Vietnamese (1)	[109]
214	1EF1	ự	LATIN SMALL LETTER U WITH HORN AND DOT BELOW	Vietnamese (1)	[109]
215	1EF3	ỳ	LATIN SMALL LETTER Y WITH GRAVE	Vietnamese (1)	[109]

#	Unicode	Glyph	Unicode Name	Languages using the code point (EGIDS)	Reference supporting inclusion (URL etc.)
216	1EF5	ỵ	LATIN SMALL LETTER Y WITH DOT BELOW	Vietnamese (1)	[109]
217	1EF7	ỷ	LATIN SMALL LETTER Y WITH HOOK ABOVE	Vietnamese (1)	[109]
218	1EF9	ỹ	LATIN SMALL LETTER Y WITH TILDE	Vietnamese (1) Guarani (1)	[109], [142]

The sequence “ss” is redundant from the repertoire point of view (it’s the same as two adjacent “s” code points) and it is not included in this table. It is however, required for proper definitions of variants. See Section 6.1.5 and Appendix D.5 for why it has to be added of the LGR. Because it can be part of valid labels, it is not outside the repertoire, bringing the total count of repertoire elements to 219.

5.3.1. Note on Combining Marks

There are seven Unicode code points included in the Latin repertoire which are non-spacing combining marks, and which are presented below in Table 4. They are not listed individually in the repertoire, since they cannot be used independently. Also, they cannot be arbitrarily combined with just any other code points from the repertoire. They are used only in specific combinations that are included as sequences in the repertoire above. (See Section 5.2.1, Inclusion Principle #3.)

Table 4. Combining Marks Included in the Repertoire of Latin Script LGR.

Unicode	Glyph	Unicode name
0300	◌̀	Combining Grave Accent
0301	◌́	Combining Acute Accent
0303	◌̃	Combining Tilde
0304	◌̄	Combining Macron
0308	◌̈	Combining Diaresis
0327	◌̧	Combining Cedilla
0331	◌̱	Combining Macron below

5.3.2. Note on Caron with Letters d, l, and t

It was raised that the following code points could be confused with the base character followed by apostrophe (U+02BC), see glyphs and fonts in Appendix D.8:

- ď U+010F Latin Small Letter D with Caron
- ľ U+013E Latin Small Letter L with Caron
- ť U+0165 Latin Small Letter T with Caron

However, as the apostrophe (U+02BC) is not included in the repertoire, there is no possibility for this confusion. The three letters are included in the repertoire as these are used in the Czech and Slovak languages.

5.4. Excluded Code Points

The Internet Architecture Board (IAB) has mandated that punctuation marks cannot be used in domain names[5]. This includes punctuation marks themselves, code points that look like punctuation marks, and letters which, although they are single letters in a particular language's alphabet, *look like* punctuation marks. It also includes cases where a diacritic is placed so that it looks like a separate punctuation mark. Accordingly, the following letters from various languages using the Latin script have been excluded from the repertoire.

Table 5. Punctuation Marks or Punctuation Mark Look-Alikes

Unicode	Glyph	Unicode Name	Language	Reference
02BB	ʻ	Modifier Letter Turned Comma	Hawaiian (2)	[135]
02BC	ʼ	Modifier Letter Apostrophe	Chamorro - (1) Dagaare-Burkina Faso (4) Dagbani (Dagomba) (4) Dholuo (5) Garo (2) Hausa (2) Mossi (3)	[140], [148], [189], [261], [262], [147], [212], [201], [264],

[5] https://www.iab.org/documents/correspondence-reports-documents/2012-2/iab-statement-the-interpretation-of-rules-in-the-icann-gtld-applicant-guidebook/

Unicode	Glyph	Unicode Name	Language	Reference
			Tartar (2) Tausūg (3) Tongan (1) Uzbek (1)	[134], [265]
A78C	'	Latin Small Letter Saltillo	Central Sinama (4) Guarani (1) Kaqchikel (4) Oromo (Afaan) (5) Pangasinan (3)	[267], [268]. [142], [143], [126], [269], [270]
01C3	!	Latin Letter Retroflex Click	Khoekhoe (4)	[235], [271], [145],[274]

Table 6. Letters Combined with Punctuation Marks or Punctuation Mark Look-Alikes.

Unicode	Glyph	Unicode Name	Language	Reference
0063 + 0068 + A78C	ch'	Latin Small Letter C + Latin Small Letter H + Latin Small Letter Saltillo	Quechua (3)	[225]
0067 + 02BC	g’	Latin Small Letter G + Modifier Letter Apostrophe	Uzbek (1)	[266]
02BC + 0068	’h	Latin Modifier Letter Apostrophe with Latin Small Letter H	Dagaare - Burkina Faso (4)	[148]
006B + A78C	k'	Latin Small Letter K + Latin Small Letter Saltillo	Quechua (3)	[225]
02BC + 006C	’l	Latin Modifier Letter Apostrophe with Latin Small Letter L	Dagaare - Burkina Faso (4)	[148]
006C + 02BC	l’	Latin Small Letter L + Modifier Letter Apostrophe	Garo (2)	[262]
006D + 02BC	m’	Latin Small Letter M + Modifier Letter Apostrophe	Garo (2)	[262]

Unicode	Glyph	Unicode Name	Language	Reference
006E + 02BC	n’	Latin Small Letter N + Modifier Letter Apostrophe	Garo (2)	[262]
006E + 0067 + 02BC	ng’	Latin Small Letter N + Latin Small Letter G + Modifier Letter Apostrophe	Garo (2)	[262]
014B + 02BC	ŋ’	Latin Small letter Eng with Modifier Letter Apostrophe	Adzera (4)	[192]
006F + 02BC	o’	Latin Small Letter O + Modifier Letter Apostrophe	Uzbek (1)	[266]
0070 + A78C	p'	Latin Small Letter O + Latin Small Letter Saltillo	Quechua (3)	[225]
0071 + A78C	q'	Latin Small Letter Q + Latin Small Letter Saltillo	Quechua (3)	[225]
0074 + A78C	t'	Latin Small Letter T + Latin Small Letter Saltillo	Quechua (3)	[225]
02BC + 0077	’w	Latin Modifier Letter Apostrophe with Latin Small Letter W	Dagaare - Burkina Faso (4)	[148]

5.4.1. Other Excluded Letters

Complete explanation for the exclusions could be found in https://www.icann.org/en/system/files/files/msr-5-overview-24jun21-en.pdf
- Section 5.7.5 (pg. 27). This is a quote from https://www.icann.org/en/system/files/files/msr-5-overview-24jun21-en.pdf
- Section 5.7.5 (pg. 27).

> *“The Integration Panel recognizes that several of these code points, in particular the following six, are widely used and prominently occur in their respective writing systems. Nevertheless, the Integration Panel concludes that security concerns outweigh an interest in more naturally mnemonic TLDs and has removed the code points from the MSR.”*

Three of the six code points referenced by the IP are listed in Table 7.

Table 7. Glyphs which are Confusables of Punctuation Marks and are Excluded from the Repertoire of Latin Script LGR.

Unicode	Glyph	Unicode Name	Language	Reference
01C0	ǀ	Latin Letter Dental Click	Khoekhoe (4)	[235], [271], [145], [274]
01C1	ǁ	Latin Letter Lateral Click	Khoekhoe (4)	[235], [271], [145], [274]
01C2	ǂ	Latin Letter Alveolar Click	Khoekhoe (4)	[235], [271], [145], [274]

In MSR-5, it is noted as the justification for exclusion that the Latin Letter Dental Click (U+01C0) resembles a vertical line. There are a variety of other glyphs which are included in the MSR also representing essentially a Vertical Line (U+007C) -- see Section 6.3.4. Nonetheless, since it is not in the MSR it is excluded from the repertoire.

A fourth letter that the Latin GP proposed for inclusion is the Middle Dot (U+00B7). This character is an integral part of the Catalan language. Because the status of this code point under IDNA 2008 is CONTEXTO and "code points permitted by IDNA 2008 under the CONTEXTO and CONTEXTJ rules are automatically excluded" according to the RZ-LGR Procedure Section B.3.4.2, this request could not be accommodated.

Table 8. CONTEXTO and CONTEXTJ Code Points Excluded from the Repertoire of Latin Script LGR.

Unicode	Glyph	Unicode Name	Language	Reference
00B7	·	Middle Dot	Catalan (2)	[272],[273]

Marshallese orthography is apparently inconsistent in using either the dot below or the cedilla (https://en.wikipedia.org/wiki/Cedilla , https://en.wikipedia.org/wiki/Marshallese_language, and https://omniglot.com/writing/marshallese.php).

Even if the confusion between the cedilla and the dot below is accepted for Marshallese, it is clear that the usage of the dot below was the result of implementation deficiency in rendering the correct sequence and used as a temporary remedy.

Therefore, the preferred representation (l, m, n, o with cedilla) are included and l, m, n with dot below will be excluded from the repertoire (o with dot below is used by other LGR languages and is included in the repertoire).

Table 8.1. The letter l, m, n with Dot Below used in Marshallese.
Excluded from the Repertoire of Latin Script LGR.

Unicode	Glyph	Unicode Name	Language	Reference
1E37	ḷ	LATIN SMALL LETTER L WITH DOT BELOW	Marshallese (1)	[213], [214], [215], [216]
1E43	ṃ	LATIN SMALL LETTER M WITH DOT BELOW	Marshallese (1)	[213], [136], [215], [216]
1E47	ṇ	LATIN SMALL LETTER N WITH DOT BELOW	Marshallese (1)	[136], [215], [216]

6. Variants

This section discusses the definition of variants for the Latin script, the discovery methodology, and the proposed candidates.

In accordance with the Procedure, an IDN variant for the Latin Root Zone LGR is an alternate code point (or sequence of code points) that could be substituted for a code point (or sequence of code points) in a candidate label to create a variant label that is considered the "same".

6.1. Principles for In-Script Variants

For the Latin Root Zone LGR the meaning of "same" will vary slightly. Latin GP determined that there are two dimensions for sameness for the Latin script:

1. visual
2. non-visual

In addition to the above, Latin GP has reviewed other cases which may or may not fall under those categories, such as IDNA 2003 compatibility and URL underlining.

For the normative specification of the LGR, a matrix has been developed, which will indicate for any codepoint why it is considered a variant. The table below lists all reasons for positively establishing a variant relationship between two or more code points, as referenced in the XML and Appendix D. However, there have been additional factors not part of this matrix, which may have prevented the GP from establishing two or more code points as variants or removed them as such. These were not the only factors in the GP's decisions for or against inclusion of variant pairs. But they were the factor that established one or another of the variant pairs found.

Table 9. Variants Principles Matrix.

Index #	Principle	Reason	Disposition
1	Visual variant (homoglyphs)	Security	Blocked
2	Visual variant (glyphs nearly identical due to font design)	Security	Blocked
3	Visual variant (generally acceptable alternate glyphs)	Security	Blocked
4	Non-visual variant	Security	Blocked
5	IDNA 2003 Compatibility	Security/ Usability	Partially Allocatable

6.1.1. Distinguishing Visual from Non-Visual Variants

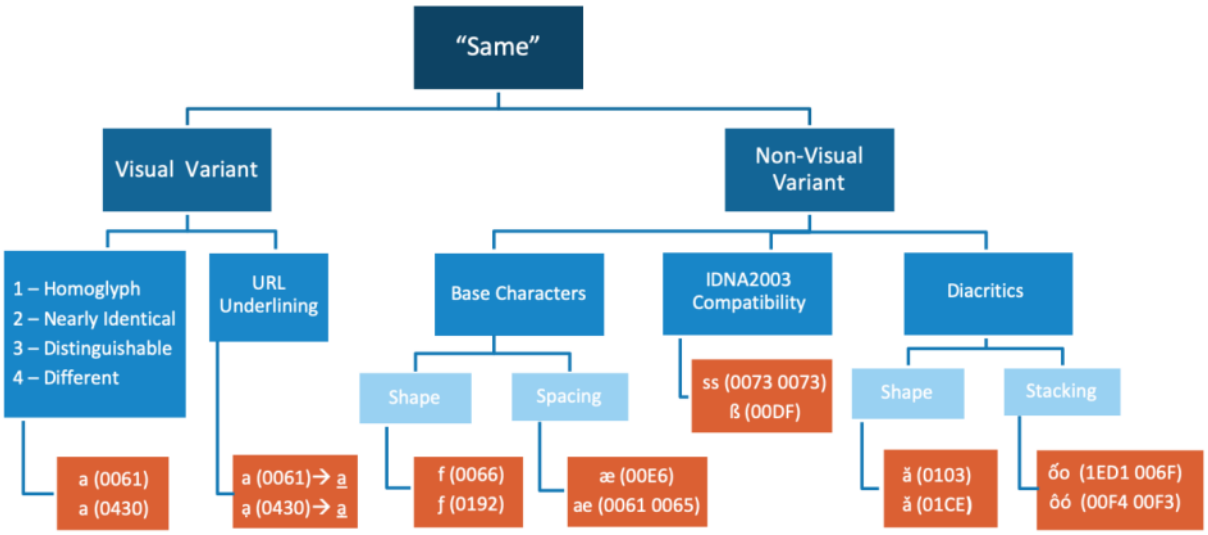
Latin GP has analyzed variants on the basis of both visual and non-visual aspects. While the criteria for visual variants are fairly consistent across both in-script and cross-script variants, the non-visual variation was less clear-cut.

With non-visual variants the issue is essentially two-fold:

1. Either readers (of domain name labels) may consider two glyphs conceptually identical, despite being able to visually tell them apart, or
2. readers may identify glyphs wrongly with other letters or sequences of letters in certain contexts.

Both issues relate to the psycholinguistic process of reading and writing. These are based not only on graphic aspects, but also on other aspects such as linguistic, contextual and cognitive factors. However, the second issue also overlaps strongly with visual equivalence. While such capacities are generally individual to single readers, Latin GP had to identify certain key areas where such non-visual equivalence may be confusable across significant parts of the script-using community and across individual readers. The GP has identified several aspects, which may influence why two or more code points may be considered “same”, as summarized in the following diagram:

Diagram 1: The Sub-Types of “Same” in Latin Script



Sections 6.1.2 and 6.1.3 below discuss first the types of visual similarity (on the left-hand branch of the diagram) then the non-visual similarity (on the right-hand branch of the diagram).

6.1.2. Visual Variants

Per [MSR], “the kinds of variants to be defined in the Root Zone LGR are limited to homoglyphs, which are characters essentially identical appearance as a result of design, instead of merely similar appearance” (22 March 2017, IP Feedback to Latin GP Proposal, Document Version 1).

However, based on discussions within the Latin GP and by the GP with Integration Panel, the panel concluded that visual similarity is not a categorical but a gradual distinction. Accordingly, Latin GP devised a four-point scale to determine whether a given pair of candidate characters tended to fall into the “essentially identical appearance as a result of design” group, i.e., clear-cut case of a homoglyph, and the “merely similar appearance” group.

This scale was found to be useful by the GP, because it places similar interpretations next to one another: While both categories Homoglyphs and Distinct vis-a-vis one another are not only self-explanatory but were also judged very coherently across different members of the GP, the debates usually revolved around the difference between a Homoglyph and Nearly Identical case, a Nearly Identical Case versus a Distinguishable case, and - to a lesser degree - a Distinguishable case versus a Distinct case. The scale thus allowed the GP to express such gradual distinctions. The levels of that scale are presented together with a concise definition below in Table 10:

Table 10. Scale for Classifying Degree of Visual Similarity

Score	Category
1	Homoglyphs A pair of code points in this category has essentially identical appearance as a result of design.
2	Nearly Identical A pair of code points is considered Nearly Identical when the visual confusion can be attributed to font design.
3	Distinguishable A pair of code points is considered Distinguishable when any of the code point's glyphs have recognizably different features from the other code point.
4	Distinct The two glyphs in the pair are sufficiently different to be distinct.

Over time, a rough consensus evolved as summarized by the concise definitions of the items in Table 10. The GP decided that a Latin code point is deemed a visual variant with another code point when the two code points or sequence of code points are either

1. homoglyphs (i.e., visual score = 1), or
2. nearly identical (i.e., visual score = 2).

Nonetheless, numerous debates took place about the precise rating between different pairs of variant candidates according to this scale. These were eventually resolved only by means of explicit rating by each active member, to establish majority decisions. However, during this very long process the GP came to the understanding that visual appearance was not the only aspect which could lead to users considering code points as variants. For pragmatic reasons, this other category, was simply termed 'Non-Visual Variant', as rendered on the right-hand branch of in Diagram 1 above, and as discussed in the following sections.

6.1.3. Non-Visual Variant: Shape of Base Characters

Historically, the classical Latin or Roman alphabet consisted of only 23 letters. Most new letters developed since are based on already existing letters and are therefore derived letters, or they were inspired by or adopted from other scripts, that is, they are borrowed letters. Derived letters were usually modified by extending certain lines (e.g., k vs. ƙ or f vs. ƒ) or by dropping elements (e.g., i vs. ɪ). In handwriting practices, a cursive writing style dominates connecting most letters to the right in order to speed up handwriting. The same kinds of changes to letters are made in order to make those connections; that is lines are extended and elements are dropped. Accordingly, Latin GP hypothesized that some hand-written forms may end up taking

similar or the same shapes as some derived letters, and that readers may consider such unknown derived letters as hand-written variations of familiar letters, such as e.g., v vs. ʋ.

Also, some letters have traditionally different shapes in hand-written and printed forms such as a vs. ɑ. (The latter shape is the traditional form encountered in handwriting. However, it is also found in some fonts. In particular, it is routinely seen when fonts are italicized) Many such differences also overlap with the difference between upper and lower case, such as e.g., e vs. ɛ, with the latter glyph being a common upper-case form in handwriting to the former glyph and letter.

6.1.4. Non-Visual Variant: Spacing of Base Characters

Several letters have been derived by putting more closely together sequences of two or more letters, and the result of such modifications of spacing in between letters are called ligatures. This strategy to develop new letters was already employed in antiquity. For example, the letter w was derived out of a sequence of two instances of the letter v, i.e., vv (https://en.wikipedia.org/wiki/History_of_the_Latin_script).

While the origins are still somewhat recognizable in the case of w, in other cases the ligatures are not recognizable anymore as combinations of their original letters, such as ß which was formed on the hand-written basis of s and z (https://en.wikipedia.org/wiki/%C3%9F). In such cases, where letters are recognizable as being composed of two or more letters, confusion could arise among readers and depending on the spacing in between those glyphs in a font (which depends on typographic factors such as e.g., kerning), ligatures may become indistinguishable from a sequence of letters of which the same ligature was originally composed.

6.1.5. Non-Visual Variant: IDNA 2003 Compatibility

In Section 5.5 of Maximal Starting Repertoire — MSR-5 Overview and Rationale, the Integration Panel highlighted risks due to IDNA compatibility issues:

“In IDNA2003, case folding is applied, which creates compatibility issues between IDNA2008 and IDNA2003 for several code points. This arguably makes the affected code points candidates for summary exclusion from the MSR on grounds of Longevity (§2.1).”

Of those code points, two belong to the Latin-script repertoire, namely 00DF Latin Small Letter Sharp S and 0131 Latin Small Letter Dotless I. The solutions based on an understanding of IDNA compatibility are presented in sections 6.4.2. The considerations involving those code points and leading to those solutions are discussed in further detail in Appendix D.5.

6.1.6. Non-Visual Variant: Shape of Diacritics

Diacritics are modifiers surrounding basic letter shapes. In some cases, diacritics are considered part and parcel of a letter shape, such as the dot on top of i. However, they are generally recognized as distinct graphic elements of the script employed to form new letters, such as é based on e featuring an acute accent on top. The majority of derived letters of Latin script were developed using this strategy.

Over time, novel diacritics became employed which were based on other diacritics. For example, ű features a base character u with a double acute (˝), a diacritic which is in turn based on the single acute (´). Many novel diacritics are very limited in use and occur in only a few languages. Typically, they were developed to express less common distinctive linguistic features of languages written in Latin script, such as Tone, and often they are only familiar to users of such languages. Essentially there are three types of potential issues with these modifiers:

Firstly, certain diacritics may be considered conceptually the same as others by some of the user community, such as cedilla below and a comma below[6], grave and hook above.

Secondly, in some cases certain diacritics are not distinct from one another in handwriting traditions, such as e.g., a caron often being written in the same way as a breve, or a dot above being written in the same way as an acute. Furthermore, in cursive hand-writing writers make use of particular strategies to write letters more quickly, modifying them in ways in which the diacritics become visually identical or confusable with others, such as a diaeresis being replaced by two vertical strokes, which could be mistaken for a double acute in italic fonts, or a tilde being written ‘simply’ as a simple horizontal stroke above, i.e., a macron. Users thus perceive the two as interchangeable.

Lastly, any given language uses only a small subset of the available diacritics. Especially with those diacritics used only in a very limited part of the script-using community, this may lead to confusion with significant parts of the script-using community or even the majority. For example, the horn (as used in combination with the basic letter shape “o” on 01A1 “ơ” Latin Small Letter O with Horn) could be mistaken by some readers for a misplaced acute (´). Or even an apostrophe (‘) -- for those users unaware that punctuation marks are excluded from use in IDN-labels because of the LDH principle.

In summary, diacritics which are different in Unicode were deemed interchangeable, or even indistinguishable, by some members of the Panel.

6.1.7. Non-Visual Variant: Stacking of Diacritics

[6] https://en.wikipedia.org/wiki/Cedilla

Diacritics are also combined with one another, such as “ắ” (1EA5, Latin Small Letter A with Circumflex and Acute) featuring both a circumflex and an acute. Such combinations are for the most part comparatively recent innovations, which again were often developed for linguistically distinctive features absent from European languages and therefore not traditionally represented in Latin script, such as tone. These novel elements of the script were often encoded in later revisions of Unicode and glyphs have been developed only for a very limited number of fonts.

In consequence, many fonts use fallback rendering, replacing missing glyphs by taking them from any other font featuring the missing glyph and available to the user’s client. In other fonts, such glyphs may be represented with overlapping or misplacement of diacritics occurring frequently. Therefore, glyphs featuring base characters with several diacritics may become visually identical or confusable to readers with sequences of glyphs featuring the same diacritics on two separate code points or may even become effectively invisible in context by crossing over into adjacent glyphs.

6.2. Methodology for Developing Variants

6.2.1. In-Script Variants

The variant situation in the Latin script is made more complex by the existence of multiple common fonts. For example, there are several common fonts which use serifs (for example Times New Roman and Courier New), and others (for example Calibri and Arial) which are san-serif. Users of the Latin script have been trained to ignore those serifs. And thus users may also ignore diacritic marks which resemble a serif. For example, in Latin Small Letter K with Hook (Unicode 0199 ƙ) the hook is on the same order as the serifs in the Latin Small Letter K alone (k) – see Appendix E.5.7. Also, letters can completely change shape between fonts. For example, the Latin Small Letter G can appear as either g (Calibri) or `g (Courier New)`, depending on the font. And finally, letters can undergo significant changes in shape with an italicized font is use. For example, Latin Small Letter A can transform from a (Normal, Times New Roman) to *a (Italic, Time News Roman)*. The former is totally unlike the Greek Letter Alpha, while the latter can readily be confused for it – see Section 6.3.2.3.

In the case of visual variants, the following cases will be proposed as in-script variant:

1. Homoglyphs (i.e., visual score = 1): when any given pair of code points or code point sequences are visually identical as represented in a common use font (e.g., Arial, Times New Roman or Courier New) by Internet applications, such as Internet browsers.
2. Nearly Identical (i.e., visual score = 2): when any given pair of code points or code point sequences are close enough visually that at least 5 of the 7 GP members could not distinguish them in at least one of these fonts.

In the case of non-visual variants, the methodology is different, and depends on the type of suspected variance:

To test the hypotheses regarding the influence of handwriting on font design and the conception of readers, Latin GP looked at both handwriting samples as well as font design. The Latin GP looked comprehensively at font design when evaluating possible variants. In addition, in some cases, Latin GP looked at how handwriting typically renders letters in order to understand other ways that users might be accustomed to visualizing particular cases. This was not done systematically, just as an aid to guide the GP's review in particular cases. In the case of glyph shape for base characters and diacritics, it was assumed that if such handwriting practices would cross-over into the printed forms, there should be fonts in which such potential variant pairs would turn out to be identical or nearly identical in appearance by a significant number of fonts.

In the case of cross-script variants, the GP initially examined glyphs only in three widely used fonts, namely Arial, Courier New, and Times New Roman. However, in the case of in-script variants the GP chose to also compare glyphs across a wide number of fonts to see if a significant minority of fonts indicated a possible variant relationship between code points. This approach was chosen because there is no stability for the fonts employed by software. Not only are different fonts used across different types of software as well as across different platforms, but most clients offer the option to change the fonts, and some protocols also allow the server to freely specify a different font.

Therefore, the only way to predict what will be a plausible case for a variant relationship, is to look for trends in the rendering of certain glyphs, and to see if a significant minority of fonts render the same glyph in a distinctly different manner. Since font designers are free to play with shapes and graphic elements comprising the glyphs recognizable by most users for a specific letter, there will always be 'extreme' cases, which may not be representative of the typical rendering of a character. However, if several fonts make use of the same graphical features in rendering of a glyph, such a shared feature may lead to a perceived similarity, which can pose a risk to stability and which may have to be dealt with at LGR-level.

In some potential variants identified by the Latin GP , a significant minority of glyphs share some features, which suggested a variant relationship to other code points. Latin GP decided that this phenomenon did not rise to the level of variant status based on a discussion among members actively participating in that discussion. In such cases the GP decided that these cases should be listed as Latin In-Script Confusables (see Appendix E). This should highlight potential risks for any party looking to implement the LGR.

The GP initially used the website https://wordmark.it/ to compare strings across a large number of different fonts. In order to attain results which were less dependent on pre-installed fonts on specific platforms and user interfaces, renderings were compared using Google Fonts, a font library employed by many APIs, instead of system fonts as rendered by the same website.

Where the shape of base characters or diacritics was suspected to lead to variant candidates, strings containing the two code points, such as ff or vice versa, i.e., U+0066 U+0192, or strings

containing code points featuring the two diacritics, such as ăă or vice versa, i.e., U+0103 U+01CE, were compared.

Sometimes spacing of base characters or stacking of diacritics were suspected to lead to variant candidates. In these cases, strings containing the ligature plus the separate elements of the ligature, such as œ and oe or vice versa (i.e., U+0153 U+006F U+0065) were compared. In other cases, GP compared strings containing code points featuring the stacked diacritics followed by the base character which the stacked diacritics modifies as well as sequences of code points featuring those diacritics separately (where available), such as e.g., ố o ô ó, i.e., U+1ED1 U+006F U+00F4 U+00F3.

This analysis was conducted for all code points featured in the suggested repertoire, as well as relevant candidates from other scripts. Eventually, the GP went through and examined each pair individually. See Appendix D.3.13.

Because the amount of difference between similar glyphs is a continuum, the decisions made on assigning variants are necessarily matters of judgement. (Appendix D.3.13 presents the GP's analysis of these.)

In marginal cases, the level of confusion (in root zone labels) may outweigh the desire for distinctive contrast (in a specific language). In general, however, the GP has assumed that the desire of distinctive contrast outweighs the concerns about confusion in root zone labels. But one case where the GP did take the view that probable overall confusion was more important is the Vietnamese contrast between A WITH GRAVE (à) and A WITH HOOK ABOVE (ả U+00E0 with U+1EA3). This is seen in section 6.3.1.3 below, and variant set 23.

Variants based on compatibility with IDNA 2003 are discussed separately below in section 6.4.2.

6.2.2. Cross-Script Variants

Latin GP has analyzed variant relationships across related scripts, such as Cyrillic, Armenian and Greek. In addition, cases where a character shape is so generic that it occurs in multiple unrelated scripts were examined. Examples of such generic shapes include a straight vertical line (Latin Small Letter L), a circle (Latin Small Letter O), and a crescent (Latin Small Letter C and Latin Small Letter Open O).

The shapes of glyphs can differ among fonts. Accordingly, Latin GP selected three fonts to represent Latin script, which it deemed to be widespread enough to be representative: Arial, Courier New, and Times New Roman. In the case of Armenian script, it was noted that there were varying glyph shapes, depending on the application used for rendering strings,

which made the initial analysis much more difficult[7]. The Latin GP consulted the Armenian Proposal to identify which glyphs the Armenian GP had chosen for representation in its Proposal [ARMENIAN] and considered those as standard for purposes of comparison with Latin script. To demonstrate the glyphs as seen and considered by Latin GP, screenshots in parts of this document are used to ensure that the reader sees the same shapes that Latin GP looked at during the analysis.

6.3. Variant Sets

6.3.1. In-Script Variants

In the following, the variant sets confirmed by Latin GP are presented together with the relevant data and rationale. The full list of potential variant pair candidates shortlisted and analyzed by the GP, including such cases which were not confirmed, is presented in Appendix D.

6.3.1.1. Variant Pairs with Diacritics: Breve and Caron

The Breve diacritic consists of a smooth curve, whereas the Caron diacritic consists of two straight lines meeting at a shallow angle. When the underlying letter is large enough, these are readily distinguishable (see Appendix D.3.1 and D.3.13). But at a normal font size (e.g., 12-point type) they are indistinguishable. Accordingly, the GP has identified the following pair of variants because the glyphs are either homoglyphs or nearly identical.

Breve				Caron			
Source Unicode Name	Source Code Point	Source Glyph	Mapping	Target Glyph	Target Code Point	Target Unicode Name	Type
Latin Small Letter G with Breve	011F	ğ	↔	ǧ	01E7	Latin Small Letter G with Caron	Blocked

[7] Google Sheets, the tool used for cross-script analysis, did not offer a variety of font designs for Armenian letters, which made it difficult for the Latin GP to replicate Armenian GP's results. Therefore, Latin GP used an alternate application, Microsoft Excel, which did offer more variety of font styles as seen in the snapshot.

6.3.1.2. Variant Pairs with Diacritics: Tilde and Macron

The Tilde diacritic consists of a wavy horizontal line whereas the Macron diacritic consists of a straight horizontal line. When the underlying letter is large enough, these are readily distinguishable (see Appendix D.3.2 and D.3.13). But at a normal font size (e.g., 12-point type) they are indistinguishable. Accordingly, GP has identified the following pairs of variants because the glyphs are either homoglyphs or nearly identical.

Tilde			Mapping	Macron			Type
Source Unicode Name	Source Code Point	Source Glyph		Target Glyph	Target Code Point	Target Unicode Name	
Latin Small Letter A with Tilde	00E3	ã	↔	ā	0101	Latin Small Letter A with Macron	Blocked
Latin Small Letter E with Tilde	1EBD	ẽ	↔	ē	0113	Latin Small Letter E with Macron	Blocked
Latin Small Letter G with Combining Tilde	0067 + 0303	g̃	↔	ḡ	1E21	Latin Small Letter G with Combining Macron	Blocked
Latin Small Letter I with Tilde	0129	ĩ	↔	ī	012B	Latin Small Letter I with Macron	Blocked
Latin Small Letter N with Tilde	00F1	ñ	↔	n̄	006E + 0304	Latin Small Letter N with Combining Macron	Blocked
Latin Small Letter O with Tilde	00F5	õ	↔	ō	014D	Latin Small Letter O with Macron	Blocked
Latin Small Letter U with Tilde	0169	ũ	↔	ū	016B	Latin Small Letter U with Macron	Blocked

6.3.1.3. Variant Pairs with Diacritics: Grave and Hook Above

The GP has identified the following pairs of variants. (see Appendix D.3.13) because the glyphs are either homoglyphs or nearly identical.

Grave			Mapping	Hook Above			Type
Source Unicode Name	Source Code Point	Source Glyph		Target Glyph	Target Code Point	Target Unicode Name	
Latin Small Letter A with Grave	00E0	à	↔	ả	1EA3	Latin Small Letter A with Hook Above	Blocked
Latin Small Letter O with Grave	00F2	ò	↔	ỏ	1ECF	Latin Small Letter O with Hook Above	Blocked
Latin Small Letter U with Grave	00F9	ù	↔	ủ	1EE7	Latin Small Letter U with Hook Above	Blocked
Latin Small Letter Y with Grave	1EF3	ỳ	↔	ỷ	1EF7	Latin Small Letter Y with Hook Above	Blocked

6.3.1.4. Variant Pairs with Diacritics: Acute and Dot Above

The GP has identified the following pairs of variants (See Appendix D.3.5 and Appendix 3.13) because the glyphs are either homoglyphs or nearly identical. However, while these consonants are variants, vowels involving the same two diacritics are merely Confusable.

Acute			Mapping	Dot Above			Type
Source Unicode Name	Source Code Point	Source Glyph		Target Glyph	Target Code Point	Target Unicode Name	
Latin Small Letter C with Acute	0107	ć	↔	ċ	010B	Latin Small Letter C with Dot Above	Blocked
Latin Small Letter N with Acute	0144	ń	↔	ṅ	1E45	Latin Small Letter N with Dot Above	Blocked
Latin Small Letter Z with Acute	017A	ź	↔	ż	017C	Latin Small Letter Z with Dot Above	Blocked

6.3.1.5. Variant Pairs with Diacritics: Acute and Hook Above

GP has identified the following pairs of variants (See Appendix D.3.13) because the glyphs are either homoglyphs or nearly identical.

Acute			Mapping	Hook Above			Type
Source Unicode Name	Source Code Point	Source Glyph		Target Glyph	Target Code Point	Target Unicode Name	
Latin Small Letter Y with Acute	00FD	ý	↔	ỷ	1EF7	Latin Small Letter Y with Hook Above	Blocked

6.3.1.6. Additional In-script Variant Pairs

GP has identified the following pairs of variants for a variety of reasons as indicated in the table below (see also Appendix D.3.13).

Source Unicode Name	Source Code Point	Source Glyph	Mapping	Target Glyph	Target Code Point	Target Unicode Name	Type	Rationale
Latin Small Letter F	0066	f	↔	ƒ	0192	Latin Small Letter F with Hook	Blocked	Generally acceptable alternate glyph See Appendix D.1.1
Latin Small Letter I	0069	i	↔	ỉ	1EC9	Latin Small Letter I with Hook Above	Blocked	Glyphs either homoglyph or nearly identical. See Appendix D.3.13
Latin Small Letter S + Latin Small Letter S	0073 0073	ss	↔	ß	00DF	Latin Small Letter Sharp S	Blocked Allocatable	IDNA 2003 Compatibility See Appendix D.5.1

Source Unicode Name	Source Code Point	Source Glyph	Mapping	Target Glyph	Target Code Point	Target Unicode Name	Type	Rationale
Latin Small Letter G with Dot Above	0121	ġ	↔	ģ	0123	Latin Small Letter G with Cedilla	Blocked	Glyphs either homoglyph or nearly identical. See Appendix D.3.13
Latin Small Letter Dotless I	0131	ı	↔	i	0069	Latin Small Letter I	Allocatable Blocked	See Appendix D.5.2
Latin Small Letter Dotless I	0131	ı	↔	ɩ	0269	Latin Small Letter Iota	Blocked	Glyphs either homoglyph or nearly identical. See Appendix D.3.13
Latin Small Letter Turned E	01DD	ə	↔	ə	0259	Latin Small Letter Schwa	Blocked	Glyphs either homoglyph or nearly identical. See Appendix D.1.14

6.3.2. Cross-Script Variants

6.3.2.1. Armenian Script

Latin GP proposes the following cross-script variants with the Armenian script. (The two tables below display the same information; the second table, however, is a screenshot taken from Microsoft Excel to demonstrate the glyph shapes as seen by the Latin GP during the cross-script variant analysis.) The details can be found in Appendix D.9.3.

Table 11. Armenian Cross-Script Variants

Source Unicode Name	Source Code Point	Source Glyph	Mapping	Target Glyph	Target Code Point	Target Unicode Name	Type	Rationale
Latin Small Letter G	0067	g	↔	g	0581	Armenian Small Letter Co	Blocked	Glyphs nearly identical due to font design
Latin Small Letter H	0068	h	↔	h	0570	Armenian Small Letter Ho	Blocked	Glyphs nearly identical due to font design
Latin Small Letter N	006E	n	↔	n	0578	Armenian Small Letter Vo	Blocked	Glyphs nearly identical due to font design
Latin Small Letter O	006F	o	↔	o	0585	Armenian Small Letter Oh	Blocked	Homoglyph
Latin Small Letter Q	0071	q	↔	q	0566	Armenian Small Letter Za	Blocked	Glyphs nearly identical due to font design
Latin Small Letter U	0075	u	↔	u	057D	Armenian Small Letter Seh	Blocked	Glyphs nearly identical due to font design
Latin Small Letter Iota	0269	ɩ	↔	ʟ	0582	Armenian Small Letter Yiwn	Blocked	Glyphs nearly identical due to font design

Screenshot taken from Microsoft Excel. The three glyphs for each code point are set in Times New Roman, Arial, and Courier, respectively:

Latin			Armenian			Disposition	Rationale
Unicode Name	Unicode	Glyph	Glyph	Unicode	Unicode Name		
LATIN SMALL LETTER O	006F	o o o	o o o	0585	ARMENIAN SMALL LETTER OH	Blocked	Homoglyph
LATIN SMALL LETTER Q	0071	q q q	q q q	0566	ARMENIAN SMALL LETTER ZA	Blocked	Glyphs nearly identical due to font design
LATIN SMALL LETTER H	0068	h h h	h h h	0570	ARMENIAN SMALL LETTER HO	Blocked	Glyphs nearly identical due to font design
LATIN SMALL LETTER N	006E	n n n	n n n	0578	ARMENIAN SMALL LETTER VO	Blocked	Glyphs nearly identical due to font design
LATIN SMALL LETTER U	0075	u u u	u u u	057D	ARMENIAN SMALL LETTER SEH	Blocked	Glyphs nearly identical due to font design
LATIN SMALL LETTER G	0067	g g g	g g g	0581	ARMENIAN SMALL LETTER CO	Blocked	Glyphs nearly identical due to font design
LATIN SMALL LETTER IOTA	0269	ɩ ɩ ɩ	ʟ ʟ ʟ	0582	ARMENIAN SMALL LETTER YIWN	Blocked	Glyphs nearly identical due to font design

6.3.2.2. Cyrillic Script

The Latin GP proposes the following cross-script variants with the Cyrillic script. The details can be found in Appendix D.9.4

Table 12: Cyrillic Cross-Script Variants

Source Unicode Name	Source Code Point	Source Glyph	Mapping	Target Glyph	Target Code Point	Target Unicode Name	Type	Rationale
Latin Small Letter R	0072	r	↔	r	0433	Cyrillic Small Letter Ghe	Blocked	Glyphs nearly identical due to font design
Latin Small Letter Y	0079	y	↔	y	04AF	Cyrillic Small Letter Straight U	Blocked	Glyphs nearly identical due to font design. (Appendix D.9.1.1)

Source Unicode Name	Source Code Point	Source Glyph	Mapping	Target Glyph	Target Code Point	Target Unicode Name	Type	Rationale
Latin Small Letter C with Cedilla	00E7	ç	↔	ç	04AB	Cyrillic Small Letter Es with Descender	Blocked	Glyphs nearly identical due to font design
Latin Small Letter Y with Diaeresis	00FF	ÿ	↔	ӳ	04F1	Cyrillic Small Letter U with Diaeresis	Blocked	Glyphs nearly identical due to font design
Latin Small Letter R with Acute	0155	ŕ	↔	ѓ	0453	Cyrillic Small Letter Gje	Blocked	Glyphs nearly identical due to font design
Latin Small Letter R with Stroke	024D	ɍ	↔	ҕ	0493	Cyrillic Small Letter Ghe with Stroke	Blocked	Glyphs nearly identical due to font design
Latin Small Letter U with Dot Below	1EE5	ụ	↔	џ	045F	Cyrillic Small Letter Dzhe	Blocked	Glyphs nearly identical due to font design. (Appendix D.9.1.2.).
Latin Small Letter A	0061	a	↔	а	0430	Cyrillic Small Letter A	Blocked	Homoglyph
Latin Small Letter C	0063	c	↔	с	0441	Cyrillic Small Letter Es	Blocked	Homoglyph
Latin Small Letter E	0065	e	↔	е	0435	Cyrillic Small Letter Ie	Blocked	Homoglyph
Latin Small Letter H	0068	h	↔	һ	04BB	Cyrillic Small Letter Shha	Blocked	Homoglyph

Source Unicode Name	Source Code Point	Source Glyph	Mapping	Target Glyph	Target Code Point	Target Unicode Name	Type	Rationale
Latin Small Letter I	0069	i	↔	i	0456	Cyrillic Small Letter Belarusian-Ukrainian I	Blocked	Homoglyph
Latin Small Letter J	006A	j	↔	j	0458	Cyrillic Small Letter Je	Blocked	Homoglyph
Latin Small Letter L	006C	l	↔	I	04CF	Cyrillic Small Letter Palochka	Blocked	Homoglyph
Latin Small Letter O	006F	o	↔	o	043E	Cyrillic Small Letter O	Blocked	Homoglyph
Latin Small Letter P	0070	p	↔	p	0440	Cyrillic Small Letter Er	Blocked	Homoglyph
Latin Small Letter S	0073	s	↔	s	0455	Cyrillic Small Letter Dze	Blocked	Homoglyph
Latin Small Letter X	0078	x	↔	x	0445	Cyrillic Small Letter Ha	Blocked	Homoglyph
Latin Small Letter Y	0079	y	↔	y	0443	Cyrillic Small Letter U	Blocked	Homoglyph
Latin Small Letter A with Diaeresis	00E4	ä	↔	ä	04D3	Cyrillic Small Letter A with Diaeresis	Blocked	Homoglyph
Latin Small Letter Ae	00E6	æ	↔	æ	04D5	Cyrillic Small Ligature A Ie	Blocked	Homoglyph

Source Unicode Name	Source Code Point	Source Glyph	Mapping	Target Glyph	Target Code Point	Target Unicode Name	Type	Rationale
Latin Small Letter E with Diaeresis	00EB	ë	↔	ё	0451	Cyrillic Small Letter Io	Blocked	Homoglyph
Latin Small Letter I with Diaeresis	00EF	ï	↔	ї	0457	Cyrillic Small Letter Yi	Blocked	Homoglyph
Latin Small Letter O with Diaeresis	00F6	ö	↔	ӧ	04E7	Cyrillic Small Letter O with Diaeresis	Blocked	Homoglyph
Latin Small Letter A with Breve	0103	ă	↔	ӑ	04D1	Cyrillic Small Letter A with Breve	Blocked	Homoglyph
Latin Small Letter H with Stroke	0127	ħ	↔	ћ	045B	Cyrillic Small Letter Tshe	Blocked	Homoglyph
Latin Small Letter Turned E	01DD	ǝ	↔	ә	04D9	Cyrillic Small Letter Schwa	Blocked	Homoglyph
Latin Small Letter Schwa	0259	ə	↔	ә	04D9	Cyrillic Small Letter Schwa	Blocked	Homoglyph
Latin Small Letter Ezh	0292	ʒ	↔	ӡ	04E1	Cyrillic Small Letter Abkhasian Dze	Blocked	Homoglyph

6.3.2.3. Greek Script

The Latin GP proposes the following cross-script variants with Greek script. The details can be found in Appendix D.9.5

Table 13: Greek Cross-Script Variants

Source Unicode Name	Source Code Point	Source Glyph	Mapping	Target Glyph	Target Code Point	Target Unicode Name	Type	Rationale
Latin Small Letter O	006F	o	↔	ο	03BF	Greek Small Letter Omicron	Blocked	Homoglyph
Latin Small Letter I with Acute	00ED	í	↔	ί	03AF	Greek Small Letter Iota with Tonos	Blocked	Homoglyph
Latin Small Letter I with Diaeresis	00EF	ï	↔	ϊ	03CA	Greek Small Letter Iota with Dialytika	Blocked	Homoglyph
Latin Small Letter O with Acute	00F3	ó	↔	ό	03CC	Greek Small Letter Omicron with Tonos	Blocked	Homoglyph
Latin Small Letter Dotless I	0131	ı	↔	ι	03B9	Greek Small Letter Iota	Blocked	Homoglyph
Latin Small Letter Open E	025B	ɛ	↔	ε	03B5	Greek Small Letter Epsilon	Blocked	Homoglyph
Latin Small Letter Iota	0269	ɩ	↔	ι	03B9	Greek Small Letter Iota	Blocked	Homoglyph

Latin Small Letter V	0076	v	↔	ν	03BD	Greek Small Letter Nu	Blocked	Glyphs nearly identical due to font design.
Latin Small Letter A	0061	a (*a*)	↔	α	03B1	Greek Small Letter Alpha	Blocked	Glyphs nearly identical due to font design. See (Appendix D.9.2.1.).
Latin Small Letter P	0070	p	↔	ρ	03C1	Greek Small Letter Rho	Blocked	Glyphs nearly identical due to font design. (Appendix D.9.2.2.)
Latin Small Letter U	0075	u	↔	υ	03C5	Greek Small Letter Upsilon	Blocked	Glyphs nearly identical due to font design. (Appendix D.9.2.3.)
Latin Small Letter Y	0079	y	↔	γ	03B3	Greek Small Letter Gamma	Blocked	Glyphs nearly identical due to font design
Latin Small Letter Sharp S	00DF	ß	↔	β	03B2	Greek Small Letter Beta	Blocked	Glyphs nearly identical due to font design. (Appendix D.9.2.4.).
Latin Small Letter A with Acute	00E1	á	↔	ά	03AC	Greek Small Letter Alpha with Tonos	Blocked	Glyphs nearly identical due to font design
Latin Small Letter U with Acute	00FA	ú	↔	ύ	03CD	Greek Small Letter Upsilon with Tonos	Blocked	Glyphs nearly identical due to font design. (Appendix D.9.2.3.).
Latin Small Letter U with Diaeresis	00FC	ü	↔	ϋ	03CB	Greek Small Letter Upsilon with Dialytika	Blocked	Glyphs nearly identical due to font design

Latin Small Letter O with Horn	01A1	ơ	↔	σ	03C3	Greek Small Letter Sigma	Blocked	Glyphs nearly identical due to font design. (Appendix D.9.2.5.)
Latin Small Letter Gamma	0263	ɣ	↔	γ	03B3	Greek Small Letter Gamma	Blocked	Glyphs nearly identical due to font design (D.9.2.6)
Latin Small Letter V with Hook	028B	ʋ	↔	υ	03C5	Greek Small Letter Upsilon	Blocked	Glyphs nearly identical due to font design. (Appendix D.9.2.3)

In addition, there may be more variants introduced by the Greek LGR proposal during the integration process. Latin GP has inspected the tentative list in the Greek LGR proposal and would accept such variants imposed on the Latin script through the integration process.

6.3.2.4. Generic Glyphs

In MSR, the Integration Panel highlights the risk of "a number of homoglyphs of code points that cross scripts", providing examples of "circle glyph" from seven scripts (See Appendix D.7):

"Because simple glyph shapes like this give effectively no hint of script identity, the IP encourages Generation Panels to consider cross-script variants in such cases even for otherwise unrelated scripts. Among related scripts, there may be pairs of code points that are identical or nearly identical despite having more complex shapes. Where these can be used to form a label that is a homograph of a label in another script, they should be investigated for variant status." [MSR, page 22-23]

Most scripts have used similar graphic elements to distinguish basic letter shapes. Accordingly, there are a few shapes which are sufficiently generic that they occur in both related and unrelated scripts[8], such as the "circle glyph" referenced by IP. For Latin script, in addition to a circle shape (Latin Small Letter O 006F) this includes a single vertical straight line (Latin Small Letter L 006C and Latin Small Letter Dotless I 0131) or a crescent (Latin Small Letter C 0053 and

[8] Only very few script creations occurred in complete isolation (cf. [DANIELS], inter alia), and most scripts have inspired one another through linguistic and cultural contact in terms of features expressed and graphic elements employed, irrespective of whether such scripts were related historically in a linguistic sense or not.

Latin Small Letter Open O 0254). While these examples are independent code points in Latin script, in other scripts they may occur as combining mark code points.

Latin GP has identified the following variant relationships based on an analysis of generic glyphs of scripts included in [MSR]. (Note that generic glyphs which were already included above for Armenian, Cyrillic, and Greek are not listed again here.) All shortlisted variant candidates which were found to be merely Confusables are presented in Appendix E.

Source Unicode Name	Source Code Point	Source Glyph	Mapping	Target Glyph	Target Code Point	Target Unicode Name	Type	Rationale
Latin Small Letter Dotless I	0131	ı	↔	ו	05D5	Hebrew Letter Vav	Blocked	Glyphs nearly identical due to font design
Latin Small Letter O	006F	o	↔	ס	05E1	Hebrew Letter Samekh	Blocked	Glyphs nearly identical due to font design
Latin Small Letter O	006F	o	↔	ഠ	0D20	Malayalam Letter Ttha	Blocked	Glyphs nearly identical due to font design
Latin Small Letter O	006F	o	↔	ဝ	101D	Myanmar Letter Wa	Blocked	Glyphs nearly identical due to font design
Latin Small Letter C	0063	c	↔	င	1004	Myanmar Letter Nga	Blocked	Glyphs nearly identical due to font design
Latin Small Letter S	0073	s	↔	ട	0D1F	Malayalam Letter Tta	Blocked	Glyphs nearly identical due to font design

6.4. Other Considerations for Variant Analysis

Latin GP has also considered two other potential security risks, which could affect the safety and stability of the root zone, namely the effect of URL underlining and full compliance with IDNA 2003 but not IDNA 2008 The results of that analysis are summarized in this section, with details of the analysis presented in Appendix D.

6.4.1. URL Underlining

Background of the issue

In true printed material italic and bold face have been used for emphasizing longer or shorter text. Typewriters, lacking those features, instead used underlining. Underlining has been unproblematic when the entire character stays above the baseline. If part of the character is below the baseline, there is a risk that the underlining hides features. Since the beginning of the web era, underlining has been routinely used to indicate a hyperlink, even though it is not mandatory to use underlining. Nowadays, many websites opt to use other methods to signal the user of the existence of a URL link, e.g., bold text. But use of underlining is still widespread.

A hyperlink in the context of HTML code consists of two parts, the URL to point at the other resource and the descriptor displayed. There is no mandatory connection between the hyperlinked URL and the descriptor displayed. The descriptor displayed can be identical to the hyperlinked URL, but it is important to know that they could differ.

The only connection between the descriptor displayed and the URL is that when the descriptor string is clicked on (or in some other way activated) the URL for that link string is activated. It is possible to have a descriptor "ICANN organization" hyperlinked to the URL "https://www.icann.org/"[9] (or "https://icann.org/"), in which case descriptor and hyperlinked URL would be the same[10]. It is equally valid, however, to have a descriptor "https://icann.org/" hyperlinked to the URL "https://iana.org/"[11], which – on click or other form of activation of the hyperlink – would resolve to "iana.org" despite having clicked on a descriptor which says "icann.org".

When links are found in an HTML document, which could be documents on the web, an email message in HTML format or even an HTML document on the local computer, for example, the links are created when the document is created. At creation, both descriptor and URL are set. As demonstrated above, there is not much to prevent misleading links from being created, in the sense that the user (whoever clicks or activates the hyperlink) draws the wrong conclusion of which URL the hyperlink should reference to. In the example above, the user would think that the descriptor would cause his web browser to open the website icann.org but instead iana.org will be opened by the web browser. If the purpose is to mislead, the descriptor and the hyperlinked URL can be chosen specifically to make it harder to see the difference than in this example, i.e., the link text "icann.org" can point at a URL going somewhere else.

Hyperlinks are not only found in HTML documents, but also in numerous other file types and documents, such as DOCX, ODT or PDF documents, with all rich-text based formats sharing the capacity to have a distinction between a descriptor and a hyperlink. Additionally, underlining is a styling preference in such document formats, but only common because users are accustomed to links to be underlined.

---

[9] E.g. in HTML: `<a href="https://icann.org">ICANN organisation</a>`
[10] E.g. in HTML: `<a href="https://icann.org">https://icann.org</a>`
[11] E.g. in HTML: `<a href="https://iana.org">https://icann.org</a>`

A third type of application where links are commonly found are email clients, where automatic creation of links to of URLs in plaintext messages is the norm. In those cases, the hyperlink was not encoded in the message sent, but it is created by the email client when it parses the message body looking for a string which can readily be interpreted as a domain name or URL (such parsers also create false positives, linkifying text strings which are simply interrupted by a dot).

The domain name is identified based on the outer shape, such as starting with a protocol identifier like “http://” or starting with “www.” Or ending in a recognized top-level domain name (such as “.com” or “.org”), and there is no common standard on which prefixes or suffixes are recognized by the parser since the behavior is dependent on the application and platform. If the text found was a complete URL (including a protocol specifying prefix), there is effectively no distinction between the descriptor and the hyperlinked URL created by the parser, if the automatic creation of links is used by the application.

If, however, a domain name without a protocol specifying prefix was identified by the parser, a URL is created from the domain name assuming the protocol to be Hypertext by prefixing “http://” to the beginning of the hyperlink. As noted above, for other file types and document formats (such as DOCX, PDF etc.) there is a choice on the side of the creator of the file or document to make the descriptor something else (or the same) as the hyperlinked URL. But email clients (and other applications, where a parser automatically linkifies plain text to hypertext) probably use underlining to indicate the link.

As was stated in the beginning of this section, underlining can hide parts of characters below the line. In text which makes use of ASCII-only characters, this is usually not a big problem even for strings with parts of the glyph cross the baseline and those parts become overlaid by the underline and therefore obfuscated, since users are trained by experience to unconsciously infer the right character when reading.

Here the focus should be on the domain name, and when the discussion is about the URL, the part after the domain name should be disregarded. Traditionally, before IDNA entered the stage, a domain name was created from ASCII letters a-z, in lower or upper case, the digits 0-9, hyphen “-” and the dot “.”. None of those can be confused even if everything below the baseline is hidden. IDNA has, however, changed that. The Latin GP has focused on the characters in the Latin script. It could be seen that there are character pairs, such as “a” vs “a̲”[12], that potentially can be confused when underlined, “a̲” vs “a̲”, depending on type face and rendering (program displaying).

In an email of August 29, 2018, the Integration Panel (IP)[13] highlighted security risks based on the underlining of labels in URLs. The IP asked the GP to take such risks into particular consideration:

> *“There are recent and widely published examples of phishing attacks using Latin IDNs in which the key features involved were diacritics below the letter. […] Of all diacritics, diacritics below can be difficult to distinguish or be prone to clipping –*

[12] LATIN SMALL LETTER A + COMBINING MACRON BELOW

[13] Reference to what it is

there is less space below the baseline than between the typical lowercase glyph and the top of the line. […] The IP would like to encourage the Latin GP (and any other GP facing cases like this) to explicitly examine this example and other cases like it, where code points can become indistinguishable in common usage scenarios for IDNs, and formally conclude whether and how to take these into account when designing their LGR."

Analysis method and Data

Based on the background discussed above, GP started to analyze all potentially confusing pairs of characters from the repertoire of the Latin script selected for the root zone.

The GP used the same methodology and framework used for the analysis of cross-script variants (see section 6.2 above). See Appendix D.6 for the data analyzed.

Underlined character pairs were compared. Underlining here is not a modification by some kind of "mark" but the text feature in the application. Example of such pairs are the already mentioned LATIN SMALL LETTER A "a" vs. the sequence LATIN SMALL LETTER A + COMBINING MACRON BELOW "a̲".

At the end of that study Latin GP found that there are actually two underlining methods. One is the "traditional" crude underlining which could be described as the result of taking a ruler and a pen and drawing a line without regarding what kind of character is underlined. The other is a more sophisticated method where the pen is lifted just before it hits something that goes below the base line and then starts again just after leaving a little space before and after (i.e., text-decoration-skip-ink)[14].

Whereas the "crude" method seems to hide some diacritics below the baseline, at least sometimes, the sophisticated method does not seem to have that problem. Which method is used is decided by the application, possibly in combination with the operating system. It is not in the hands of the user.

To see what the issue might be, consider, for example,

xaxa̲xa̲xa̲xbxb̲xcxç̲xdxd̲xd̲̲xexe̲xe̲xexe̲xixi̲xj̲xjxkxk̲xyxy̲xy̲x

vs.

xaxa̲xa̲xa̲xbxb̲xcxç̲xdxd̲xd̲̲xexe̲xe̲xexe̲xixi̲xj̲xjxkxk̲xyxy̲xy̲x

Conclusion

The GP concludes that underlining can create confusion and make otherwise distinct glyphs indistinguishable, which could be an issue in a domain name context. However, GP's conclusion is also that creating variants of such cases will not resolve the issue of spoofing since in most cases, the descriptor displayed is in the hand of the creator and can be connected to any URL and domain name. Latin GP has not designated any variants due to underlining.

[14] https://developer.mozilla.org/en-US/docs/Web/CSS/text-decoration-skip-ink

6.4.2. IDNA 2003 Compatibility

The Latin GP has analyzed and discussed the pros and cons of different solutions for mitigating risks arising from IDNA 2003 compatibility issues, as discussed in detail in Appendix D.5.

6.4.2.1. Latin Small Letter Sharp S

In the case of Latin Small Letter Sharp S (00DF), the preliminary detailed analysis is presented in Appendix D.5.1. Latin GP proposes a solution which includes the code point together with a variant relationship with the sequence of letters ‘ss’ (0073 0073), as follows:

Table 16. In-Script Variants for Latin Small Letter Sharp S (00DF)

Source Unicode Name	Source Code Point	Source Glyph	Mapping	Target Glyph	Target Code Point	Target Unicode Name	Type
Latin Small Letter Sharp S	00DF	ß	→	ss	0073 0073	Latin Small Letter S + Latin Small Letter S	Allocatable
Latin Small Letter S + Latin Small Letter S	0073 0073	ss	→	ß	00DF	Latin Small Letter Sharp S	Blocked

1. As these code point variants are also alternate forms, these are made allocatable type (in one direction). “ß” and “ss” can co-occur within the same German domain labels, therefore it is plausible for a label to contain the two sequences;
2. It is common for users in Germany to use a “ß” in a label in some cases, and never replace it with “ss”, while
3. German speaking users in Switzerland would prefer a label with “ss” in all cases.

Since “ss” and “s” coexist in the repertoire, and “s” has variant relationships on its own, these variants overlap. There is a need to explicitly determine all variant relationships to ensure the entire variant set is well-behaved for index variant calculation. The sequence "ss" can also have variants. If variants occur, when Sharp S is replaced by “ss” or a variant of it, it must be replaced in all cases. For example, the two sequences: U+0455 U+0455 (ss), U+0D1F U+0D1F (SS).

It is desirable to minimize the number of allocatable label variants. Accordingly, the Panel decided that, if at least one occurrence of Sharp S is replaced, all of the occurrences of Sharp S must be replaced as well.

6.4.2.2. Latin Small Letter Dotless I

The GP decided the Latin Small Letter Dotless I (0131) and the Latin Small Letter I (0069) are variants. The detailed analysis is presented in Appendix D.5.2.

Source Unicode Name	Source Code Point	Source Glyph	Mapping	Target Glyph	Target Code Point	Target Unicode Name	Type
Latin Small Letter I	0069	i	→	ı	0131	Latin Small Letter Dotless I	Blocked
Latin Small Letter Dotless I	0131	ı	→	i	0069	Latin Small Letter I	Allocatable

Usually Latin Capital Letter I (0049) is the upper case of Latin Small Letter I (0069), and Latin Small Letter I is the lower case of Latin Capital Letter I. At the same time, Latin Capital Letter I is also the upper case of Latin Small Letter Dotless I (0131). For Latin Small Letter Dotless I, the case relationship is therefore asymmetrical.

However, in the settings for two languages (so called “system [locale] settings”), Turkish and Azeri, the case relationship is different. In those settings only, Latin Small Letter I and Latin Capital Letter I with Dot Above (0130) are in a mutual upcase/downcase relationship to each other, and Latin Small Letter Dotless I (0131) and Latin Capital Letter I in another. These special case behavior of Latin Capital Letter I with Dot Above and Latin Small Letter Dotless I, respectively, only applies to the Turkish system locale settings and Azeri system locale settings.

Applications, e.g. most browsers, down case any Latin script non-ASCII string before IDNA conversion, therefore because the down casing is locale dependent in the case described above there is a risk of misdirection, e.g., a Turkish user types ‘BUS.EXI’ in the browser bar thinking of the website ‘bus.exı’, but the browser resolves to ‘bus.exi’, i.e., using the dominating case relationship instead of the Turkish case relationship. To be on the safe side the Latin GP has decided to make Latin Small Letter I and Latin Small Letter Dotless I variants of each other in this proposal for Latin script.

As noted above, if a label contains multiple Sharp S letters, the only allocatable variant label allowed must change all of them to “ss”. Similarly, with a Dotless I.

However, in the event that a label contains both multiple Sharp Ss, and multiple Dotless Is, then there are three allocatable variant labels:

1. One with all occurrences of the Sharp S changed, but all of occurrences of the Dotless I retained.
2. One with all of the occurrences of the Dotless I changed, but all of the occurrences of the Sharp S retained.
3. One with all occurrences of the Sharp S change, and all of the occurrences of the Dotless I changes as well.

In-Script Variant Mapping Types

Specialized variant mappings have been defined to limit the allocable variant labels with Sharp S and Dotless I as discussed above.

U+00DF (ß) Sharp S has been given the reflexive variant type "r-eszett" and U+0131 (ı) Dotless i has been given the reflexive variant of type "r-dotless".

The variant mapping from U+00DF (ß) Sharp S is to "ss" is of type "eszett-to-ss", while the variant type for the mapping from "ss" to Sharp S is "blocked".

The variant mapping from U+0131 (ı) Dotless i to "i" is of type "dotted", while the variant type for the mapping from "i" to Dotless i is "blocked".

Details can be found in the XML.

6.5. Variant Due to Transitivity

Transitivity is a mathematical property of relations, defined as follows: if two mathematical expressions are in relation to a third expression then both are in the same relation to each other. The equality relation is a good example having the transitivity property (if A = B and B = C, then A = C).

The variant relation is functionally a “same-as” relation. The variant, when substituted in place of the original, should result in a label that is perceived as the “same” by human readers. Human perceptions do not work like mathematics. For labels that are look-alikes, there is a continuous transition from precisely identical appearance to mere similarity. But there are also other dimensions of “sameness” that may occur for labels in general, from identical meaning to identical pronunciation. The rule can give a result that two glyphs end up being variants of each other due to transitivity which, if compared directly, would not be candidates for variants.

Another problem arises, when variants are not based on having identical appearance but merely being deceptively close. Especially when cross-script variants are involved, transitivity may produce variants that are unacceptable. For example, the work of the Latin, Cyrillic, and Greek GPs resulted in a chain of variants involving multiple glyphs. Transitivity then resulted in the Latin Small Letter V (v) as an in-script variant of the Latin Small Letter Y (y). But these are both ASCII characters which, by rule, cannot be variants of each other. Accordingly, the Latin, Greek, and Cyrillic GPs had to sit down and negotiate which link in the chain would be broken. That is, which variant relationship would be reduced to "merely Confusable", something that does not require transitivity and also cannot be captured in a variant definition.

6.6. Additional Discussion on Variants

It was suggested to the Latin GP that, in the Fula language [149], Latin Small Letter N with Tilde (U+00F1, ñ) is used interchangeably with Latin Small Letter N with Left Hook (U+0272, ɲ). However, the GP was unable to find definitive confirmation for this. Accordingly, we have not included these two code points as variants.

6.7. Complete Variant Sets

Based on the discussion on variants above. There are 50 variant sets in Latin LGR including 16 in-script variant sets and 34 cross-script variant sets. The complete variant sets are shown in Table 17.

Table 17 Complete variant sets in Latin LGR

#	Latin Letter	Armenian Letter	Cyrillic Letter	Greek Letter	Other Script Letter
1	a U+0061 LATIN Small Letter A á U+00E1 Latin Small Letter A with Acute (Imposed in-script variant by Greek LGR)		a U+0430 Cyrillic Small Letter A	α U+03B1 Greek Small Letter Alpha ά U+03AC Greek Small Letter Alpha with Tonos	

#	Latin Letter	Armenian Letter	Cyrillic Letter	Greek Letter	Other Script Letter
2	c U+0063 Latin Small Letter C		c U+0441 Cyrillic Small Letter Es		c U+1004 Myanmar Letter Nga
3	e U+0065 Latin Small Letter E		e U+0435 Cyrillic Small Letter Ie		
4	f U+0066 Latin Small Letter F ƒ U+0192 Latin Small Letter F with Hook				
5	g U+0067 Latin Small Letter G	g U+0581 Armenian Small Letter Co			
6	g̃ U+0067 U+0303 Latin Small Letter G with Combining Tilde ḡ U+1E21 Latin Small Letter G with Combining Macron				
7	h U+0068 Latin Small Letter H	h 0570 Armenian Small Letter Ho	h U+04BB Cyrillic Small Letter Shha		

#	Latin Letter	Armenian Letter	Cyrillic Letter	Greek Letter	Other Script Letter
8	i U+0069 Latin Small Letter I í U+00ED Latin Small Letter I with Acute ï U+00EF Latin Small Letter I with Diaeresis ı U+0131 Latin Small Letter Dotless I ỉ U+1EC9 Latin Small Letter I with Hook Above ɩ U+0269 Latin Small Letter Iota (Some are imposed in-script variant by Greek LGR)	ւ U+0582 Armenian Small Letter Yiwn	і U+0456 Cyrillic Small Letter Byelorussian-Ukrainian I ї U+0457 Cyrillic Small Letter Yi	ί U+03AF Greek Small Letter Iota with Tonos ι U+03B9 Greek Small Letter Iota ϊ U+03CA Greek Small Letter Iota with Dialytika ΐ U+0390 Greek Small Letter Iota with Dialytika and Tonos	ו U+05D5 Hebrew Letter Vav
9	j U+006A Latin Small Letter J		ј U+0458 Cyrillic Small Letter Je		

#	Latin Letter	Armenian Letter	Cyrillic Letter	Greek Letter	Other Script Letter
10	l U+006C Latin Small Letter L		I U+04CF Cyrillic Small Letter Palochka		
11	n U+006E Latin Small Letter N ŋ U+014B Latin Small Letter Eng ń U+0144 Latin Small Letter N with Acute ṅ U+1E45 Latin Small Letter N with Dot Above (Some are imposed in-script variant by Greek LGR)	ŋ U+0572 Armenian Small Letter GHAD n U+0578 Armenian Small Letter VO		η U+03B7 Greek Small Letter Eta ή U+03AE Greek Small Letter ETA with Tonos	
12	n̄ U+006E +U+0304Latin Small Letter N with Combining Macron ñ U+00F1 Latin Small Letter N with Tilde				

#	Latin Letter	Armenian Letter	Cyrillic Letter	Greek Letter	Other Script Letter
13	o U+006F Latin Small Letter O ó U+00F3 Latin Small Letter O with Acute (Imposed in-script variant by Greek LGR)	օ U+0585 Armenian Small Letter Oh	о U+043E Cyrillic Small Letter O	ο U+03BF Greek Small Letter Omicron ό U+03CC Greek Small Letter Omicron with Tonos	ס U+05E1 Hebrew Letter Samekh ଠ U+0B20 Oriya Letter Ttha ഠ U+0D20 Malayalam Letter Ttha ဝ U+101D Myanmar Letter Wa
14	p U+0070 Latin Small Letter P		р U+0440 Cyrillic Small Letter Er	ρ U+03C1 Greek Small Letter Rho	
15	q U+0071 Latin Small Letter Q	զ U+0566 Armenian Small Letter Za			
16	r U+0072 Latin Small Letter R		г U+0433 Cyrillic Small Letter Ghe		
17	s U+0073 Latin Small Letter S		ѕ U+0455 Cyrillic Small Letter Dze		S U+0D1F Malayalam Letter Tta

#	Latin Letter	Armenian Letter	Cyrillic Letter	Greek Letter	Other Script Letter
18	ss U+0073 U+0073 Latin Small Letter S Latin Small Letter S ß U+00DF Latin Small Letter Sharp S		ѕs U+0455 U+0455 Cyrillic Small Letter Dze Cyrillic Small Letter Dze	β U+03B2 Greek Small Letter Beta	SS U+0D1F U+0D1F Malayalam Letter Tta Malayalam Letter Tta
19	u U+0075 Latin Small Letter U ú U+00FA Latin Small Letter U with Acute ü U+00FC Latin Small Letter U with Diaeresis ʊ U+028B Latin Small Letter V with Hook (Some are imposed in-script variant by Greek LGR)	u U+057D Armenian Small Letter Seh		υ U+03C5 Greek Small Letter Upsilon ύ U+03CD Greek Small Letter Upsilon with Tonos ϋ U+03CB Greek Small Letter Upsilon with Dialytika ΰ U+03B0 Greek Small Letter Upsilon with Dialytika and Tonos	
20	v U+0076 Latin Small Letter V			ν U+03BD Greek Small Letter Nu	
21	x U+0078 Latin Small Letter X		х U+0445 Cyrillic Small Letter Ha		

#	Latin Letter	Armenian Letter	Cyrillic Letter	Greek Letter	Other Script Letter
22	ɣ U+0079 Latin Small Letter Y ɣ U+0263 Latin Small Letter Gamma		у U+0443 Cyrillic Small Letter U ү U+04AF Cyrillic Small Letter Straight U	γ U+03B3 Greek Small Letter Gamma	
23	à U+00E0 Latin Small Letter A with Grave ả U+1EA3 Latin Small Letter A with Hook Above				
24	ã U+00E3 Latin Small Letter A with Tilde ā U+0101 Latin Small Letter A with Macron				
25	ä U+00E4 Latin Small Letter A with Diaeresis		ӓ U+04D3 Cyrillic Small Letter A with Diaeresis		
26	æ U+00E6 Latin Small Letter Ae		ӕ U+04D5 Cyrillic Small Ligature A Ie		

#	Latin Letter	Armenian Letter	Cyrillic Letter	Greek Letter	Other Script Letter
27	ç U+00E7 Latin Small Letter C with Cedilla		ç U+04AB Cyrillic Small Letter Es with Descender		
28	ë U+00EB Latin Small Letter E with Diaeresis		ë U+0451 Cyrillic Small Letter Io		
29	ò U+00F2 Latin Small Letter O with Grave ỏ U+1ECF Latin Small Letter O with Hook Above				
30	õ U+00F5 Latin Small Letter O with Tilde ō U+014D Latin Small Letter O with Macron				
31	ö U+00F6 Latin Small Letter O with Diaeresis		ö U+04E7 Cyrillic Small Letter O with Diaeresis		

#	Latin Letter	Armenian Letter	Cyrillic Letter	Greek Letter	Other Script Letter
32	ù U+00F9 Latin Small Letter U with Grave ủ U+1EE7 Latin Small Letter U with Hook Above				
33	ý U+00FD Latin Small Letter Y with Acute ỳ U+1EF3 Latin Small Letter Y with Grave ỷ U+1EF7 Latin Small Letter Y with Hook Above				
34	ÿ U+00FF Latin Small Letter Y with Diaeresis		ӳ U+04F1 Cyrillic Small Letter U with Diaeresis		
35	ă U+0103 Latin Small Letter A with Breve		ӑ U+04D1 Cyrillic Small Letter A with Breve		

#	Latin Letter	Armenian Letter	Cyrillic Letter	Greek Letter	Other Script Letter
36	ć U+0107 Latin Small Letter C with Acute ċ U+010B Latin Small Letter C with Dot Above				
37	ē U+0113 Latin Small Letter E with Macron ẽ U+1EBD Latin Small Letter E with Tilde				
38	ğ U+011F Latin Small Letter G with Breve ǧ U+01E7 Latin Small Letter G with Caron				
39	ġ U+0121 Latin Small Letter G with Dot Above ģ U+0123 Latin Small Letter G with Cedilla				

#	Latin Letter	Armenian Letter	Cyrillic Letter	Greek Letter	Other Script Letter
40	ħ U+0127 Latin Small Letter H with Stroke		ħ U+045B Cyrillic Small Letter Tshe		
41	ĩ U+0129 Latin Small Letter I with Tilde ī U+012B Latin Small Letter I with Macron				
42	ŕ U+0155 Latin Small Letter R with Acute		ѓ 0453 Cyrillic Small Letter Gje		
43	ũ U+0169 Latin Small Letter U with Tilde ū U+016B Latin Small Letter U with Macron				
44	ź U+017A Latin Small Letter Z with Acute ż U+017C Latin Small Letter Z with Dot Above				

#	Latin Letter	Armenian Letter	Cyrillic Letter	Greek Letter	Other Script Letter
45	ơ U+01A1 Latin Small Letter O with Horn			σ U+03C3 Greek Small Letter Sigma ς U+03C2 Greek Small Letter Final Sigma	
46	ə U+01DD Latin Small Letter Turned E ə U+0259 Latin Small Letter Schwa		ә U+04D9 Cyrillic Small Letter Schwa		
47	ɍ U+024D Latin Small Letter R with Stroke		ҍ U+0493 Cyrillic Small Letter Ghe with Stroke		
48	ɛ U+025B Latin Small Letter Open E			ε U+03B5 Greek Small Letter Epsilon έ U+03AD Greek Small Letter EPSILON with Tonos	

#	Latin Letter	Armenian Letter	Cyrillic Letter	Greek Letter	Other Script Letter
49	ʒ U+0292 Latin Small Letter Ezh		ӟ U+04E1 Cyrillic Small Letter Abkhasian Dze		
50	ụ U+1EE5 Latin Small Letter U with Dot Below		џ U+045F Cyrillic Small Letter Dzhe		

7. Whole Label Evaluation Rules (WLE) and contextual Rules

In LGR contextual rules or restrictions can be defined in several ways. One technique is called Whole Label Evaluation Rules (WLE).

For Latin LGR no WLEs are planned. The only code points that need contextual restrictions are the non-spacing marks (see section 5.3.1). The restriction on those is that they are only allowed, in the Latin LGR, after specific letter code points. That restriction is achieved by not listing the marks as individual code points in the LGR, but only as part of the permitted sequence of a letter code point and the non-space mark (or, the sequence of a letter code point plus two ordered non-space marks).

For Latin-specific actions assigning dispositions to variant labels see Section 6.4.2.

8. Contributors

Bill Jouris
Chris Dillon (Chair of Latin GP until 2016)
Dennis Tan Tanaka
Hazem Hezzah
Jean Paul Nkurunziza
Mats Dufberg
Meikal Mumin
Michael Bauland
Mirjana Tasić (Chair of Latin GP from 2016)

ICANN Staff:
Sarmad Hussain

Pitinan Kooarmornpatana

9. References

Dates represent access dates unless indicated otherwise.

[99], The Unicode Consortium, Unicode® 11.0.0, http://www.unicode.org/versions/Unicode11.0.0/, 5 September 2018
[100], ICANN, Second Level Reference Label Generation Rules for Spanish, https://www.icann.org/sites/default/files/packages/lgr/lgr-second-level-spanish-30aug16-en.html, 31 August 2018
[101], Omniglot, Czech (čeština), http://www.omniglot.com/writing/czech.htm , 31 August 2018
[102], Omniglot, Icelandic (Íslenska), http://www.omniglot.com/writing/icelandic.htm , 31 August 2018
[103], Omniglot, Faroese (føroyskt mál), http://www.omniglot.com/writing/faroese.htm , 31 August 2018
[105], Omniglot, Chuukese (Chuuk), http://www.omniglot.com/writing/chuukese.htm , 31 August 2018
[106], SCRIPTSOURCE, Galician written with Latin script, http://www.webcitation.org/6siTI8ieQ , 31 August 2018
[107], Omniglot, Lule Sámi (julevsámegiella), http://www.omniglot.com/writing/lulesami.htm , 31 August 2018
[108], Wikipedia, Northern Sami, https://en.wikipedia.org/wiki/Northern_Sami , 4 September 2018
[109], Omniglot, Vietnamese (tiếng việt / 喃越), http://www.omniglot.com/writing/vietnamese.htm , 4 September 2018
[110], Omniglot, Romanian (limba română), http://www.omniglot.com/writing/romanian.htm , 4 September 2018
[113], Omniglot, Skolt Sámi (Sää´mǩiõll / Nuõrttsää'm), http://www.omniglot.com/writing/skoltsami.htm , 4 September 2018
[114], Omniglot, French (français), http://omniglot.com/writing/french.htm , 4 September 2018
[115], Omniglot, West Frisian (Frysk), http://www.omniglot.com/writing/westfrisian.htm , 4 September 2018
[116], Omniglot, Friulian (furlan/marilenghe), http://www.omniglot.com/writing/friulian.htm , 4 September 2018
[117], SIL International, Pequeno dicionário: Xavante-Português, Português-Xavante, https://www.sil.org/resources/archives/17019 , 1 October 2020
[119], Omniglot, German (Deutsch), http://www.omniglot.com/writing/german.htm , 4 September 2018
[120], Omniglot, Finnish (suomi), http://www.omniglot.com/writing/finnish.htm , 4 September 2018

[121], Omniglot, Turkmen (Türkmen dili / Түркмен дили), http://www.omniglot.com/writing/turkmen.htm , 4 September 2018
[122], Omniglot, Estonian (eesti keel), http://www.omniglot.com/writing/estonian.htm , 4 September 2018
[123], Omniglot, Swedish (svenska), http://www.omniglot.com/writing/swedish.htm , 4 September 2018
[124], Omniglot, Yapese (Waab), http://www.omniglot.com/writing/yapese.htm , 4 September 2018
[125], Omniglot, Dinka (Thuɔŋjäŋ), https://www.omniglot.com/writing/dinka.php , 4 September 2018
[126], Omniglot, Kaqchikel (Kaqchikel Ch'ab'äl), http://www.omniglot.com/writing/kaqchikel.htm , 4 September 2018
[127], Omniglot, Bashkir/Bashkort (Башҡорт теле / Başqort tele), http://www.omniglot.com/writing/bashkir.htm , 4 September 2018
[128], Omniglot, Alsatian (Ëlsässisch), https://www.omniglot.com/writing/alsatian.htm , 4 September 2018
[129], Wikipedia, Nuer language, https://en.wikipedia.org/wiki/Nuer_language , 4 September 2018
[130], Omniglot, Italian (italiano), http://www.omniglot.com/writing/italian.htm , 4 September 2018
[131], Wikipedia, Italian orthography, https://en.wikipedia.org/wiki/Italian_orthography , 4 September 2018
[132], Omniglot, Wolof (Wollof), http://www.omniglot.com/writing/wolof.htm , 4 September 2018
[133], Omniglot, Latvian (latviešu valoda), http://www.omniglot.com/writing/latvian.htm , 4 September 2018
[134], Omniglot, Tongan (Faka-Tonga), http://www.omniglot.com/writing/tongan.htm , 4 September 2018
[135], Omniglot, Hawaiian (ʻŌlelo Hawaiʻi), http://www.omniglot.com/writing/hawaiian.htm , 4 September 2018
[136], Omniglot, Marshallese (kajin ṃajeḷ), http://www.omniglot.com/writing/marshallese.php , 4 September 2018
[137], Omniglot, Polish (polski), http://www.omniglot.com/writing/polish.htm , 4 September 2018
[138], Omniglot, Lithuanian (lietuvių kalba), http://www.omniglot.com/writing/lithuanian.htm , 4 September 2018
[139], Omniglot, Danish (dansk), http://www.omniglot.com/writing/danish.htm , 4 September 2018
[140], Omniglot, Chamorro (chamoru), http://www.omniglot.com/writing/chamorro.htm , 4 September 2018
[141], Omniglot, Umbundu (Úmbúndú), http://www.omniglot.com/writing/umbundu.htm , 4 September 2018
[142], Omniglot, Guaraní (Avañe'ẽ), http://www.omniglot.com/writing/guarani.htm , 4 September 2018

[143], Wikipedia, Guarani alphabet, https://en.wikipedia.org/wiki/Guarani_alphabet , 4 September 2018
[144], Omniglot, Nauruan (Ekaiairũ Naoero), http://www.omniglot.com/writing/nauruan.htm , 4 September 2018
[145], Omniglot, Khoekhoe (Khoekhoegowab), https://www.omniglot.com/writing/khoekhoe.htm , 4 September 2018
[146], Omniglot, Nuer (Naath), https://www.omniglot.com/writing/nuer.htm , 4 September 2018
[147], Omniglot, Hausa (Harshen Hausa / هَرْشَن هَوْسَ), http://www.omniglot.com/writing/hausa.htm , 4 September 2018
[148], Omniglot, Dagaare, http://www.omniglot.com/writing/dagaare.htm , 4 September 2018
[149], Omniglot, Fula (Fulfulde, Pulaar, Pular'Fulaare), http://www.omniglot.com/writing/fula.htm , 4 September 2018
[150], Omniglot, Croatian (Hrvatski), http://www.omniglot.com/writing/croatian.htm , 4 September 2018
[151], Omniglot, Serbian (српски / srpski), http://www.omniglot.com/writing/serbian.htm , 4 September 2018
[152], Wikipedia, Polish language, https://en.wikipedia.org/wiki/Polish_language , 4 September 2018
[153], Omniglot, Slovak (slovenčina), http://www.omniglot.com/writing/slovak.htm , 4 September 2018
[154], Evertype Publishing, Lithuanian lietuvių kalba Version 1.1, http://www.evertype.com/alphabets/lithuanian.pdf , 4 September 2018
[157], Omniglot, Turkish (Türkçe), http://www.omniglot.com/writing/turkish.htm , 4 September 2018
[158], Omniglot, Kurdish (Kurdî / کوردی), http://www.omniglot.com/writing/kurdish.htm , 4 September 2018
[159], Omniglot, Azerbaijani (آذربایجانجا دیلی / Азәрбајҹан дили / Azərbaycan dili), http://www.omniglot.com/writing/azeri.htm , 4 September 2018
[160], Omniglot, Basque (euskara), http://www.omniglot.com/writing/basque.htm , 4 September 2018
[161], Wikipedia, Basque language, https://en.wikipedia.org/wiki/Basque_language#Writing_system , 4 September 2018
[163], Omniglot, Maltese (Malti), http://www.omniglot.com/writing/maltese.htm , 4 September 2018
[164], Omniglot, Venda (Tshivenḓa / Luvenḓa), http://www.omniglot.com/writing/venda.htm , 4 September 2018
[166], Wikipedia, Hausa language, https://en.wikipedia.org/wiki/Hausa_language , 4 September 2018
[167], Christian Chanard and Rhonda L. Hartell. 2014, Pulaar sound inventory (AA), http://phoible.org/inventories/view/809#tsource , 4 December 2019
[168], Omniglot, Brahui (Bráhuí / براوی), https://www.omniglot.com/writing/brahui.htm , 4 September 2018

[169], Wikipedia, Fon language, https://en.wikipedia.org/wiki/Fon_language , 4 September 2018
[170], Omniglot, Ewe (Eʋegbe), http://www.omniglot.com/writing/ewe.htm , 4 September 2018
[172], Omniglot, Sorbian (hornjoserbsce/dolnoserbski), https://www.omniglot.com/writing/sorbian.htm , 4 September 2018
[173], Peace corps, Botswana, An Introduction to Setswana Language, http://files.peacecorps.gov/multimedia/audio/languagelessons/botswana/Bw_Setswana_Language_Lessons.pdf , 4 September 2018
[174], Omniglot, Tswana (Setswana), http://omniglot.com/writing/tswana.php , 4 September 2018
[175], Wikipedia, Afrikaans, https://en.wikipedia.org/wiki/Afrikaans , 4 September 2018
[176], Omniglot, Albanian (shqip / gjuha shqipe), http://www.omniglot.com/writing/albanian.htm , 4 September 2018
[177], Wikipedia, Albanian alphabet, https://en.wikipedia.org/wiki/Albanian_alphabet , 4 September 2018
[179], Wikipedia, Uyghur Latin alphabet, https://en.wikipedia.org/wiki/Uyghur_Latin_alphabet , 4 September 2018
[180], Omniglot, Drehu (De'u), http://www.omniglot.com/writing/drehu.php , 4 September 2018
[182], Omniglot, Haitian Creole (Kreyòl ayisyen), http://www.omniglot.com/writing/haitiancreole.htm , 4 September 2018
[183], Wikipedia, Haitian Creole, https://en.wikipedia.org/wiki/Haitian_Creole#Orthography , 4 September 2018
[184], Omniglot, Minangkabau (Baso Minangkabau / باسو مينڠكاباو), http://www.omniglot.com/writing/minangkabau.htm , 4 September 2018
[185], Omniglot, Palauan (a tekoi er a Belau), http://www.omniglot.com/writing/palauan.htm , 4 September 2018
[186], Omniglot, Cubeo (pãmié), http://www.omniglot.com/writing/cubeo.htm , 4 September 2018
[187], Editorial Alberto Lleras Camargo, Diccionario Ilustrado Bilingüe cubeo-español español-cubeo, https://www.sil.org/system/files/reapdata/10/58/27/10582785843693992331766506069073895620/40337_01.pdf , 4 September 2018
[188], Omniglot, Inari Saami (Anarâškielâ), http://www.omniglot.com/writing/inarisami.htm , 4 September 2018
[189], Omniglot, Compiled by Wolfram Siegel, DAGBANI, http://www.omniglot.com/charts/dagbani.pdf , 4 September 2018
[190], Omniglot, Ewondo, http://www.omniglot.com/writing/ewondo.php , 4 September 2018
[191], Omniglot, Luganda (Oluganda), http://www.omniglot.com/writing/ganda.php , 4 September 2018
[192], Omniglot, Adzera, http://www.omniglot.com/writing/adzera.htm , 4 September 2018
[193], Omniglot, Ga (Gã), http://www.omniglot.com/writing/ga.htm , 4 September 2018

[194], Omniglot, Duala (Duálá), http://www.omniglot.com/writing/duala.php , 4 September 2018
[195], Omniglot, Soga (Lusoga), http://www.omniglot.com/writing/soga.htm , 4 September 2018
[196], Omniglot, Alur (Lur), http://www.omniglot.com/writing/alur.htm , 4 September 2018
[197], Omniglot, Mandinka (Mandi'nka kango / لغة مندنكا), http://www.omniglot.com/writing/mandinka.htm , 4 September 2018
[198], Omniglot, Acholi (Lwo), https://www.omniglot.com/writing/acholi.htm , 4 September 2018
[199], Omniglot, Bambara (Bamanankan), http://www.omniglot.com/writing/bambara.htm , 4 September 2018
[200], Omniglot, Raga (Hano), http://www.omniglot.com/writing/raga.htm , 4 September 2018
[201], Omniglot, Tatar (tatarça / татарча / تاتارچا), http://www.omniglot.com/writing/tatar.htm , 4 September 2018
[202], Omniglot, Zaza (Zazaki / زازاکی), https://www.omniglot.com/writing/zazaki.htm , 4 September 2018
[203], Wikipedia, Turkish alphabet, https://en.wikipedia.org/wiki/Turkish_alphabet , 4 September 2018
[204], School of English, Adam Michiewicz University, Poznań, Poland, Poznań Studies in Contemporary Linguistics 43(1),2007, pp. 169-180, A Demographic Igbo Orthography, https://www.degruyter.com/downloadpdf/j/psicl.2007.43.issue-1/v10010-007-0009-0/v10010-007-0009-0.pdf , 4 September 2018
[205], Omniglot, Igbo (Asụsụ Igbo), http://www.omniglot.com/writing/igbo.htm , 4 September 2018
[206], ItalianPod101, Italian Accents and Proper Italian Pronunciation, https://www.italianpod101.com/italian-accents , 4 September 2018
[208], Reverso Dictionary, venerdì translation | Italian-English dictionary, http://dictionary.reverso.net/italian-english/venerd%C3%AC , 4 September 2018
[209], Omniglot, Kikuyu (Gĩkũyũ), http://www.omniglot.com/writing/kikuyu.htm , 4 September 2018
[210], Omniglot, Hixkaryána, http://www.omniglot.com/writing/hixkaryana.htm , 4 September 2018
[211], Omniglot, Maasai (ɔl Maa), http://www.omniglot.com/writing/maasai.htm , 4 September 2018
[212], Omniglot, Mossi (Mòoré), http://www.omniglot.com/writing/mossi.htm , 4 September 2018
[213], Omniglot, Jenesis. The Bible in Marshallese, 2009., Contributed by Wolfgang Kuhl, http://www.omniglot.com/babel/marshallese.htm , 4 September 2018
[214], Wikipedia, Cedilla, https://en.wikipedia.org/wiki/Cedilla#Marshallese , 4 September 2018
[215], Wikipedia, Marshallese language, https://en.wikipedia.org/wiki/Marshallese_language#Display_issues , 4 September 2018
[216], Trussel, Marshallese-English Online Dictionary, http://www.trussel2.com/MOD/ , 4 September 2018

[218], Omniglot, Susu (Sosoxi), https://www.omniglot.com/writing/susu.htm , 4 September 2018
[219], Omniglot, Zarma (Zarmaciine), https://www.omniglot.com/writing/zarma.htm , 4 September 2018
[220], Omniglot, Pitjantjatjara, https://www.omniglot.com/writing/pitjantjatjara.htm , 4 September 2018
[221], Omniglot, Spanish (español/castellano), http://www.omniglot.com/writing/spanish.htm , 4 September 2018
[222], Omniglot, Filipino (wikang Filipino), http://www.omniglot.com/writing/filipino.htm , 4 September 2018
[223], Omniglot, Chavacano, http://www.omniglot.com/writing/chavacano.php , 4 September 2018
[224], Wikipedia, Ilocano language, https://en.wikipedia.org/wiki/Ilocano_language#Modern_alphabet , 4 September 2018
[225], Omniglot, Quechua (Runasimi), http://www.omniglot.com/writing/quechua.htm , 4 September 2018
[226], Wikipedia, Quechua alphabet, https://en.wikipedia.org/wiki/Quechua_alphabet , 4 September 2018
[227], Omniglot, Cape Verdean Creole (Kriolu), http://www.omniglot.com/writing/kriol.php , 4 September 2018
[228], Omniglot, Waray-Waray, http://www.omniglot.com/writing/waray.php , 4 September 2018
[229], Omniglot, Lozi (siLozi), http://www.omniglot.com/writing/lozi.htm , 4 September 2018
[230], africanlanguages.com, Sesotho sa Leboa (Northern Sotho), http://africanlanguages.com/northern_sotho/ , 4 September 2018
[231], Omniglot, Low German (Plattdüütsch / Nedderdüütsch), https://www.omniglot.com/writing/lowgerman.htm , 4 September 2018
[232], Wikipedia, Chechen language, https://en.wikipedia.org/wiki/Chechen_language , 4 September 2018
[233], Omniglot, Hungarian (magyar), http://www.omniglot.com/writing/hungarian.htm , 4 September 2018
[234], Wikipedia, Hungarian alphabet, https://en.wikipedia.org/wiki/Hungarian_alphabet , 4 September 2018
[235], Encyclopedia Britanica,Khoisan Languages, https://www.britannica.com/topic/Khoisan-languages
[236], Omniglot, Lingala, http://www.omniglot.com/writing/lingala.htm , 4 September 2018
[237], Omniglot, Akan, https://www.omniglot.com/writing/akan.htm , 4 September 2018
[238], Wikipedia, Mossi language, https://en.wikipedia.org/wiki/Mossi_language , 4 September 2018
[239], SIL-Sudan, OCCASIONAL PAPERS in the study of SUDANESE LANGUAGES No. 9, https://www.sil.org/system/files/reapdata/10/06/46/100646256099282892829790816212446104791/OPSL_9.pdf (p. 75), 4 September 2018
[240], Omniglot, Kanuri, http://www.omniglot.com/writing/kanuri.htm , 4 September 2018

[241], Omniglot, Bugis (Basa Ugi), http://www.omniglot.com/writing/bugis.htm , 4 September 2018
[242], Omniglot, Mizo (Mizo ṭawng), http://www.omniglot.com/writing/mizo.htm , 4 September 2018
[243], Omniglot, Miskito (Mískitu), http://www.omniglot.com/writing/miskito.htm , 4 September 2018
[245], Wikipedia, Papiamento, https://en.wikipedia.org/wiki/Papiamento , 4 September 2018
[246], Omniglot, Papiamento (Papiamentu), http://www.omniglot.com/writing/papiamento.php , 4 September 2018
[247], Omniglot, Chichewa (Chicheŵa), http://www.omniglot.com/writing/chichewa.php , 4 September 2018
[248], Native Languages of the Americas website, Vocabulary in Native American Languages: Mam Words, http://www.native-languages.org/mam_words.htm , 4 September 2018
[249], Omniglot, Mam (Qyol Mam), http://www.omniglot.com/writing/mam.htm , 4 September 2018
[250], Wikipedia, Pulaar language, https://en.wikipedia.org/wiki/Pulaar_language , 4 September 2018
[251], Wikipedia, Fula language, https://en.wikipedia.org/wiki/Fula_language#Writing_systems , 4 September 2018
[252], Wikipedia, Polish alphabet, https://en.wikipedia.org/wiki/Polish_alphabet , 4 September 2018
[253], Wikipedia, French orthography, https://en.wikipedia.org/wiki/French_orthography , 4 September 2018
[254], Omniglot, Yoruba (Èdè Yorùbá), https://www.omniglot.com/writing/yoruba.htm , 4 September 2018
[255], Omniglot, Esperanto, http://www.omniglot.com/writing/esperanto.htm , 4 September 2018
[256], Omniglot, Welsh (Cymraeg), http://www.omniglot.com/writing/welsh.htm , 4 September 2018
[257], Wikipedia, List of Latin-script letters, https://en.wikipedia.org/wiki/List_of_Latin-script_letters , 4 September 2018
[258], Omniglot, Montenegrin, https://www.omniglot.com/writing/montenegrin.htm, 20 March 2019
[259], Wikipedia, Rho, https://en.wikipedia.org/wiki/Rho, 24 September 2019
[261], Omniglot, Dholuo, https://www.omniglot.com/writing/dholuo.php, 4 December 2019
[262], Omniglot, Garo, https://www.omniglot.com/writing/garo.htm , 4 December 2019
[264], Omniglot, Tausug, https://www.omniglot.com/writing/tausug.htm , 4 December 2019
[265], Omniglot, Uzbek, http://www.omniglot.com/writing/uzbek.htm , 4 December 2019
[266], Wikipedia, Uzbek language, https://en.wikipedia.org/wiki/Uzbek_alphabet#Distinct_characters , 4 December 2019
[267], Omniglot, Central Sinama, https://www.omniglot.com/writing/centralsinama.htm , 4 December 2019
[268], The Central Sinama Alphabet, http://sinama.org/bahasa-sinama/sama-alphabet/ , 4 December 2019

[269], Omniglot, Oromo, https://www.omniglot.com/writing/oromo.htm , 4 December 2019
[270], Omniglot, Pangasinan, https://www.omniglot.com/writing/pangasinan.htm , 4 December 2019
[271], Wikipedia, Khoe Languages, https://en.wikipedia.org/wiki/Khoe_languages , 4 December 2019
[272], Omniglot, Catalan, http://www.omniglot.com/writing/catalan.htm , 4 December 2019
[273], Wikipedia, Interpunct, Catalan, https://en.wikipedia.org/wiki/Interpunct#Catalan , 4 December 2019
[274], Khoikhoi Language Nation https://khoekhoegowab.wordpress.com/weekly-photo-journal/text/,4 December 2019
[275], Omniglot, Shavante, https://www.omniglot.com/writing/shavante.php, 24 September 2020
[276], Malagasy Language, https://en.wikipedia.org/wiki/Malagasy_language, 24 September 2020
[277], Serer language, https://en.wikipedia.org/wiki/Serer_language , 6 April 2021
[278], Kpelle language, https://en.wikipedia.org/wiki/Kpelle_language, 6 April 2021

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

[Requirements] Integration Panel “Requirements for LGR Proposals from Generation Panels”. https://www.icann.org/en/system/files/files/Requirements-for-LGR-Proposals- 20150424.pdf

[Considerations] VIP Study Group “Considerations in the use of the Latin script in variant internationalized top-level domains” (Los Angeles, California: ICANN, October 2011). https://archive.icann.org/en/topics/new-gtlds/latin-vip-issues-report-07oct11-en.pdf

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

[Katz & Frost 1992] Katz, Leonard & Ram Frost. 1992. “The Reading Process is Different for Different Orthographies: The Orthographic Depth Hypothesis”. *Haskins Laboratories Status Report on Speech Research* 111/112. 147–160.

[Wikipedia-Latin script] Latin script. Cached version retrieved 2017-02-14. http://www.webcitation.org/6oGZwoNUu

[Wikipedia-Capital ß] Capital ß. Cached version retrieved 2018-01-17. http://www.webcitation.org/6wXlGtfqc

[Wikipedia - Ejectives] Ejectives. Cached version retrieved 2018-01-19. http://www.webcitation.org/6waqfVtj3

[Wikipedia - ASCII] ASCII. Cached version retrieved 2018-01-20. http://www.webcitation.org/6waqfVtj3

[Rogers] Rogers, Henry. 2005. *Writing systems: A linguistic approach*. Malden, Massachusetts: Blackwell Publishing.

[MSR] Maximal Starting Repertoire https://www.icann.org/resources/pages/msr-2015-06-21-en

[ARMENIAN] Armenian Generation Panel, "Proposal for an Armenian Script Root Zone LGR. Version 3." (Los Angeles, California: ICANN, June 2015. https://www.icann.org/public-comments/proposal-armenian-lgr-2015-07-22-en

[CYRILLIC] Cyrillic Generation Panel, "Proposal for Cyrillic Script Root Zone Label Generation Rules. Version 1.4." (Los Angeles, California: ICANN, October 2017. https://www.icann.org/public-comments/cyrillic-lgr-2017-10-17-en

[DANIELS] Daniels, Peter T. 1992. "The syllabic origin of writing and the segmental origin of the alphabet." *The Linguistics of Literacy* in Downing, Pamela A., Lima, Susan D., & Noonan, Michael (Eds.), 83-110. John Benjamins, Amsterdam.

[HUSSAIN] Sarmad Hussain, Ahmed Bakhat, Nabil Benamar, Meikal Mumin & Inam Ullah (2016) Enabling multilingual domain names: adfd.1.5
dressing the challenges of the Arabic script top-level domains, *Journal of Cyber Policy*, 1:1, 107-129, DOI: 10.1080/23738871.2016.1157618

[IDNA 2003] "Internationalizing Domain Names in Applications (IDNA)", RFC 3490, https://datatracker.ietf.org/doc/html/rfc3490

[IDNA 2008] "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework", RFC 5890, https://datatracker.ietf.org/doc/html/rfc5890

[Locale] System locale settings, https://en.wikipedia.org/wiki/Locale_(computer_software)