Proposal for new identifier type values

Asmus Freytag and Michel Suignard, 15 March 2022

Overview

This proposal requests the addition of two (non-exclusive) values for Identifier_Type in UTS#39 to cover some scenarios that are being distinguished in existing identifier implementations and might be useful more generally.

It is probably useful to recognize that providing these Identifier_Type values for excluded scripts (and code points) is a pointless exercise. Once could go further, and consider that rolling this out across limited_use scripts is likewise something that we could (perhaps should) formally defer at this point; therefore, the explicit examples in this proposal focuses on recommended scripts.

Confusable with Punctuation

There are a number of characters that are highly confusable or outright identical with punctuation marks from the ASCII or General punctuation block, or with other characters that could be misinterpreted generically as delimiting an identifier. These characters should not be unreservedly "recommended" for identifiers; at the minimum, it should be easy to identify them as a security risk via a dedicated property.

Therefore, we propose a new identifier_type value "punctuation_like", to be considered as characters that should be excluded by default, except if explicitly allowed.

The following discussion from the DNS Root Zone MSR document[1] provides background and a starter set:

> ... the following code points are highly confusable with or outright homoglyphs of code points, such as common punctuation characters like apostrophe or exclamation mark that are not PVALID in IDNA2008 or excluded for other reasons:
>
> - U+01C0..U+01C3 |..(!) LATIN LETTER DENTAL CLICK..LATIN LETTER RETROFLEX CLICK
> - U+02B9..U+02C1 ʹ..ˁ MODIFIER LETTER PRIME..MODIFIER LETTER REVERSED GLOTTAL STOP
> - U+02C6..U+02D1 ˆ..(˙) MODIFIER LETTER CIRCUMFLEX ACCENT..MODIFIER LETTER HALF TRIANGULAR COLON
> - U+02EC ˬ MODIFIER LETTER VOICING
> - U+02EE ˮ MODIFIER LETTER DOUBLE APOSTROPHE

[1] Integration Panel, "Maximal Starting Repertoire — MSR-5 Overview and Rationale", 24 June 2021, https://www.icann.org/en/system/files/files/msr-5-overview-24jun21-en.pdf

- U+A78C ' LATIN SMALL LETTER SALTILLO

In particular, U+01C0 and U+01C1 are indistinguishable from the punctuation marks U+007C and U+2016 in certain user interface fonts. There are other code points with glyphs that look more or less like a straight line, but their glyphs show some variation in length, width, side bearing and distance from the base line: the Integration Panel views the indistinguishable appearance as the relevant criterion in this instance. U+01C2 has a more distant resemblance to a line-drawing symbol U+256A; it is included here for consistency. U+01C3 is always indistinguishable from an exclamation mark.

Note: the MSR only considered "Recommended" scripts (excluding Bopomofo) and does not consider the digits. Limiting this identifier_type to recommended scripts would in our view be appropriate. The goal is to enforce the notion that not all identifier_status=recommended code points are fully unproblematic; the goal cannot be to trawl all ancient scripts for similar cases.

Confusable with ASCII

Coexistence of IDNs with LDH (ASCII letter, digit + hyphen) is a common scenario for zones in the DNS. The scenario is probably more common than arbitrary mixtures of scripts. The DNS shares with other identifier systems the issue that ASCII identifiers predate IDNs and/or that ASCII identifiers are far more common in some zones than IDNs. ASCII characters also include many that are simple and generic in shape, like “o”, “l”, “c” and “s” and that therefore present confusability issues across a wide range of script, not limited to European scripts derived from Greek letters.

It's also how users who are not aware that an identifier system supports non-ASCII characters can be tricked. (The scenario where non-ASCII characters are allowed, in principle, but not actually used by habit or convention is probably not uncommon).

This is a problem both for source code as well as for the DNS (for example, in Latin, the use of ASCII fallbacks remains prevalent and users may not expect alternate base shapes to represent unique letters).

Offline discussions have suggested that highlighting this subset could be useful, possibly as an identifier type that is a subset of the recommended set of identifier characters

In the RZ LGR created by ICANN for IDN TLDs, these ASCII confusables are always explicitly listed where they occur. Other cross-script mappings may be defined, but are treated as implicitly effective in the LGR specification. Rationale for the difference in treatment is that in the DNS in particular, ASCII (LDH) labels often existed in a zone before IDNs were added. Therefore collisions are more likely for these sets than for arbitrary cross-script variants.

See also the list below excerpted from the Root Zone LGR.

Note: the RZ LGR does not consider digits, so the analysis would need to be extended for the Unicode identifier types.

Excerpt from RZ-LGR

The following provides a list of variant sets that have an ASCII lowercase letter as a member. The list is an excerpt from a recent Root Zone LGR draft (the full draft should be publicly available shortly). (The count of members often includes accented versions of the Latin letter, these have been excluded here manually to emphasize the ASCII base letters).

The RZ-LGR project[2] is focused on recommended scripts (minus Bopomofo, which is considered too specialized in its field of application). Restricting the proposed Identifier_Type of ASCII confusable to the same set of scripts would not be inappropriate. It would represent a compromise between catching the most useful cases while keeping a lid on the cost of research and maintenance.

As mentioned, digits aren't considered in this particular list, but some work on an extended set has been carried out by ICANN for the "Reference LGRs for the Second Level".

Variant Set 1 — 5 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0061	a	03B1	α	↔	blocked	[105], [109], [118]	Cross-script near homoglyph
0061	a	0430	a	↔	blocked	[105], [109], [118]	Cross-script homoglyph
03B1	α	0430	a	↔	blocked	[105], [109], [118]	Cross-script near homoglyph

Variant Set 2 — 3 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0063	c	0441	c	↔	blocked	[105], [118]	Cross-script homoglyph

[2] See https://icann.org/idn and look for Root Zone LGR for the latest status and documents.

0063	c	1004		↔	blocked	[118], [120]	Cross-script near homoglyph
0441	c	1004		↔	blocked	[118]	

Note: Myanmar 105A will look like 1004 in many contexts, but not in labels that mimic an LDH label like ccc, coco etc. In other words, 105A requires the presence of one of a set of Myanmar-unique code points to make it look like 1004 or 0063. However, for confusability across Myanmar labels, 1004 and 105A clearly count. It's a known weakness of the Unicode Confusables model that it excludes sequences and contexts. (See Myanmar LGR proposal on http://icann.org/idn for details).

Variant Set 3 — 2 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0065	e	0435	e	↔	blocked	[105], [118]	Cross-script homoglyph

Variant Set 4 — 2 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0066	f	0192	ƒ	↔	blocked	[118]	Generally acceptable alternate glyph

Variant Set 5 — 2 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0067	g	0581	g	↔	blocked	[102], [118]	Cross-script near homoglyph

Variant Set 7 — 3 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0068	h	04BB		↔	blocked	[102], [105], [118]	Cross-script homoglyph
0068	h	0570	h	↔	blocked	[102], [105],	Cross-script near

						[118]	homoglyph
04BB	ɦ	0570	h	↔	blocked	[102], [105], [118]	Cross-script near homoglyph

Variant Set 8 — 14 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0069	i	0131	ı	↔	blocked	[102], [105], [109], [118]	IDNA2003 Compatibility
0069	i	0269	ɩ	↔	blocked	[102], [105], [109], [118]	Required for integration
0069	i	03B9	ι	↔	blocked	[102], [105], [109], [118]	Cross-script near homoglyph
0069	i	0456	і	↔	blocked	[102], [105], [109], [118]	Cross-script homoglyph
0069	i	0582	ւ	↔	blocked	[102], [105], [109], [118]	
0069	i	05D5	ו	↔	blocked	[112], [118]	Cross-script near homoglyph
0131	ı	0269	ɩ	↔	blocked	[102], [105], [109], [118]	Glyphs either homoglyph or nearly identical
0131	ı	03B9	ι	↔	blocked	[102], [105], [109], [118]	Cross-script homoglyph
0131	ı	0456	і	↔	blocked	[102], [105], [109], [118]	

0131	ı	0582	ʟ	↔	blocked	[102], [105], [109], [118]	
0131	ı	05D5	ן	↔	blocked	[118]	Cross-script near homoglyph
0269	ɩ	03B9	ι	↔	blocked	[102], [105], [109], [118]	Cross-script homoglyph
0269	ɩ	0456	і	↔	blocked	[102], [105], [109], [118]	
0269	ɩ	0582	ʟ	↔	blocked	[102], [105], [109], [118]	Cross-script near homoglyph
0269	ɩ	05D5	ן	↔	blocked	[118]	
03B9	ι	0456	і	↔	blocked	[102], [105], [109], [118]	Cross-script near homoglyph
03B9	ι	0582	ʟ	↔	blocked	[102], [105], [109], [118]	Cross-script near homoglyph
03B9	ι	05D5	ן	↔	blocked	[118]	
0456	і	0582	ʟ	↔	blocked	[102], [105], [109], [118]	
0456	і	05D5	ן	↔	blocked	[118]	
0582	ʟ	05D5	ן	↔	blocked	[118]	

Variant Set 9 — 2 Members

Source	Glyph	Target	Glyph	Type	Ref	Comment

006A	j	0458	j	↔	blocked	[105], [118]	Cross-script homoglyph

Variant Set 10 — 2 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
006C	l	04CF	I	↔	blocked	[105], [118]	Cross-script homoglyph

Variant Set 11 — 8 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
006E	n	014B	ŋ	↔	blocked	[102], [109], [118]	Required for integration
006E	n	03B7	η	↔	blocked	[102], [109], [118]	Cross-script near homoglyph
006E	n	0572	ɳ	↔	blocked	[102], [109], [118]	
006E	n	0578	n	↔	blocked	[102], [109], [118]	Cross-script near homoglyph
014B	ŋ	03B7	η	↔	blocked	[102], [109], [118]	Cross-script near homoglyph
014B	ŋ	0572	ɳ	↔	blocked	[102], [109], [118]	Cross-script near homoglyph
014B	ŋ	0578	n	↔	blocked	[102], [109], [118]	
03B7	η	0572	ɳ	↔	blocked	[102], [109], [118]	Cross-script near homoglyph
03B7	η	0578	n	↔	blocked	[102], [109],	Cross-script near

						[118]	homoglyph
0572	ɳ	0578	n	↔	blocked	[102], [109], [118]	Required for integration

Variant Set 12 — 10 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
006F	o	03BF	o	↔	blocked	[102], [105], [109], [118]	Cross-script homoglyph
006F	o	043E	o	↔	blocked	[102], [105], [109], [118]	Cross-script homoglyph
006F	o	0585	o	↔	blocked	[102], [105], [109], [118]	Cross-script homoglyph
006F	o	05E1	ɒ	↔	blocked	[112], [118]	Cross-script near homoglyph
006F	o	0B20	O	↔	blocked	[118], [119], [120], [121]	Cross-script near homoglyph
006F	o	0D20	O	↔	blocked	[118], [119], [120], [121]	Cross-script near homoglyph
006F	o	101D	○	↔	blocked	[118], [119], [120], [121]	Cross-script near homoglyph
03BF	o	043E	o	↔	blocked	[102], [105], [109], [118]	Cross-script homoglyph
03BF	o	0585	o	↔	blocked	[102], [105], [109], [118]	Cross-script homoglyph

03BF	o	05E1	ס	↔	blocked	[118]	
03BF	o	0B20	ଠ	↔	blocked	[118]	
03BF	o	0D20	ഠ	↔	blocked	[118]	
03BF	o	101D	ဝ	↔	blocked	[118]	
043E	o	0585	o	↔	blocked	[102], [105], [109], [118]	Cross-script homoglyph
043E	o	05E1	ס	↔	blocked	[118]	
043E	o	0B20	ଠ	↔	blocked	[118]	
043E	o	0D20	ഠ	↔	blocked	[118]	
043E	o	101D	ဝ	↔	blocked	[118]	
0585	o	05E1	ס	↔	blocked	[118]	
0585	o	0B20	ଠ	↔	blocked	[118]	
0585	o	0D20	ഠ	↔	blocked	[118]	
0585	o	101D	ဝ	↔	blocked	[118]	

05E1	ס	0B20	O	↔	blocked	[118]	
05E1	ס	0D20	O	↔	blocked	[118]	
05E1	ס	101D	○	↔	blocked	[118]	
0B20	O	0D20	O	↔	blocked	[118], [119], [120], [121]	Cross-script homoglyph
0B20	O	101D	○	↔	blocked	[118], [119], [120], [121]	Cross-script homoglyph
0D20	O	101D	○	↔	blocked	[118], [119], [120], [121]	Cross-script homoglyph

Variant Set 13 — 3 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0070	p	03C1	ρ	↔	blocked	[105], [109], [118]	Cross-script near homoglyph
0070	p	0440	p	↔	blocked	[105], [109], [118]	Cross-script homoglyph
03C1	ρ	0440	p	↔	blocked	[105], [109], [118]	Cross-script near homoglyph

Variant Set 14 — 2 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0071	q	0566	q	↔	blocked	[102], [118]	Cross-script near homoglyph

Variant Set 15 — 2 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0072	r	0433	r	↔	blocked	[105], [118]	Cross-script near homoglyph

Variant Set 16 — 3 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0073	s	0455	s	↔	blocked	[105], [118]	Cross-script homoglyph
0073	s	0D1F	ѕ	↔	blocked	[118], [119]	Cross-script near homoglyph
0455	s	0D1F	ѕ	↔	blocked	[118]	

Variant Set 17 — 5 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0073 0073	ss	00DF	ß	↔	blocked	[105], [109], [118]	IDNA2003 Compatibility
0073 0073	ss	03B2	β	↔	blocked	[105], [109], [118]	
0073 0073	ss	0455 0455	ѕѕ	↔	blocked	[105], [109], [118]	Cross-script homoglyph
0073 0073	ss	0D1F 0D1F	ѕѕ	↔	blocked	[118], [119]	Cross-script near homoglyph
00DF	ß	03B2	β	↔	blocked	[105], [109], [118]	Cross-script near homoglyph
00DF	ß	0455 0455	ss	↔	blocked	[105], [109], [118]	

00DF	ß	0D1F 0D1F	SS	↔	blocked	[118]
03B2	β	0455 0455	ss	↔	blocked	[105], [109], [118]
03B2	β	0D1F 0D1F	SS	↔	blocked	[118]
0455 0455	ss	0D1F 0D1F	SS	↔	blocked	[118]

Variant Set 18 — 9 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0075	u	028B	ʊ	↔	blocked	[102], [109], [118]	Required for integration
0075	u	03C5	υ	↔	blocked	[102], [109], [118]	Cross-script near homoglyph
0075	u	057D	u	↔	blocked	[102], [109], [118]	Cross-script near homoglyph
028B	ʊ	03C5	υ	↔	blocked	[102], [109], [118]	Cross-script near homoglyph
028B	ʊ	057D	u	↔	blocked	[102], [109], [118]	
03C5	υ	057D	u	↔	blocked	[102], [109], [118]	Cross-script near homoglyph
03CB	ϋ	057D	u	↔	blocked	[102], [109], [118]	

03CD	ύ	057D	u	↔	blocked	[102], [109], [118]	

Variant Set 19 — 2 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0076	v	03BD	ν	↔	blocked	[109], [118]	Cross-script near homoglyph

Variant Set 20 — 2 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0078	x	0445	х	↔	blocked	[105], [118]	Cross-script homoglyph

Variant Set 21 — 5 Members

Source	Glyph	Target	Glyph		Type	Ref	Comment
0079	y	0263	ɣ	↔	blocked	[105], [109], [118]	Required for integration
0079	y	03B3	γ	↔	blocked	[105], [109], [118]	Cross-script near homoglyph
0079	y	0443	у	↔	blocked	[105], [109], [118]	Cross-script homoglyph
0079	y	04AF	ү	↔	blocked	[105], [109], [118]	Cross-script near homoglyph
0263	ɣ	03B3	γ	↔	blocked	[105], [109], [118]	Cross-script near homoglyph
0263	ɣ	0443	у	↔	blocked	[105], [109], [118]	

0263	ɣ	04AF	ү	↔	blocked	[105], [109], [118]	
03B3	γ	0443	у	↔	blocked	[105], [109], [118]	Cross-script near homoglyph
03B3	γ	04AF	ү	↔	blocked	[105], [109], [118]	Cross-script near homoglyph
0443	у	04AF	ү	↔	blocked	[105], [109], [118]	Required for integration