Permitted Code Points for European Country-Code Registries

Marcel Schneider, Kim Davies
26 October 2003
Version 0.3 (Draft)

This document aims to define a broad selection of code-points, useful for deploying European languages as a whole whilst still constraining permissible code-point selection to a minimal set to reduce problems. Each code point is documented with its known applicability, so further refinement of policy can be made on a language-by-language basis.

Code-points that are not explicitly permitted by the registry, under this approach, are disallowed.

Policy by Language and Locale

Note: Only normalised forms are used throughout this part of the document. That is, input code points may include non-normalised forms (capital letters, etc.). See the table for details.

Generic Latin (LDH)

In this document, we consider the base Latin character set to consist of the 26 consecutive letters A (a_{U+0061}) through Z (z_{U+007A}), the ten consecutive digits 0 (0_{U+0030}) through 9 (9_{U+0039}), and a hyphen ($-_{U+002D}$). These codepoints are generally the permissible subset currently allowed for non-IDN registrations by registries.

German (DE)

The common German alphabet is comprised of LDH, plus the vowels A with diaresis ($ä_{U+00E4}$), O with diaresis ($ö_{U+00F6}$), and U with diaresis ($ü_{U+00FC}$). Another character in use is the the *eszett* ($ß_{U+0073}$) – but during name preparation this is convert into two latin S characters ($ss_{U+0073,U+0073}$). Note: this process can not be universally reversed – such that an “SS” can not then be returned to an eszett without appropriate context assessment.

- Latin letter A with diaeresis (ä $_{U+00E4}$)
- Latin letter O with diaeresis (ö $_{U+00F6}$)
- Latin letter U with diaeresis (ü $_{U+00FC}$)

It is common for German characters to be represented as ligatures. A with an umlaut can also be represented as AE ($ae_{(U+0061,U+0065)}$, joined as $æ_{U+00E6}$), and O with a diaresis as OE ($oe_{(U+006F,U+0065)}$, joined as $œ_{U+0153}$), and an U with a diaresis ($ue_{(U+0075,U+0065}$). This may want to be considered when generating bundles or naming rules, although these variations are not combined when using nameprep.

Source: Institüt für Deutsche Sprache

Hungarian (HU)

- LDH
- Latin letter A with acute (á U+00E1)
- Latin letter E with acute (é U+00E9)
- Latin letter I with acute (í U+00ED)
- Latin letter O with acute (ó U+00F3)
- Latin letter O with diæresis (ö U+00F6)
- Latin letter O with double acute (ő U+0151)
- Latin letter U with acute (ú U+00FA)
- Latin letter U with diaeresis (ü U+00FC)
- Latin letter U with double acute (ű U+0171)

Source: Internet Szolgáltatók Tanácsa

Icelandic (IS)

- LDH
- Latin letter A with acute (á U+00E1)
- Latin letter AE ligature (æ U+00E6)
- Latin letter E with acute (é U+00E9)
- Latin letter I with acute (í U+00ED)
- Latin letter Eth (ð U+00F0)
- Latin letter O with acute (ó U+00F3)
- Latin letter O with diæresis (ö U+00F6)
- Latin letter U with acute (ú U+00FA)
- Latin letter Y with acute (ý U+00FD)
- Latin letter thorn (þ U+00FE)

Source: ISNIC

Lithuanian (LT)

- Latin letter A with ogonek (ą U+0105)
- Latin letter C with caron (č U+010D)
- Latin letter E with dot above (ė U+0117)
- Latin letter E with ogonek (ę U+0119)
- Latin letter I with ogonek (į U+012F)
- Latin letter S with caron (š U+0161)
- Latin letter U with macron (ū U+016B)
- Latin letter U with ogonek (ų U+0173)
- Latin letter Z with caron (ž U+017E)

Source: Litnet

Maltese (MT)

- LDH
- Latin letter C with dot above (ċ U+010B)
- Latin letter G with dot above (ġ U+0121)
- Latin letter H with stroke (ħ U+0127)
- Latin letter Z with dot above (ż U+017C)

Usage notes:
The letter C (c U+0063) does not naturally occur in the language. An ie ligature is considered to be an independent unbreakable character.

Source: .mt Registry

Norway (NO)

The superset of Norwegian languages (Bokmål (NO-NB), Nynorsk (NO-NN), Northern Sami (NO-SME), Southern Sami (NO-SMA), Lule Sami (NO-SMJ)) is comprised of LDH, plus:

- Latin letter A with acute (á U+00E1)
- Latin letter A with grave (à U+00E0)
- Latin letter A with diaeresis (ä U+00E4)
- Latin letter C with caron (č U+010D)
- Latin letter C with cedilla (ç U+00E7)
- Latin letter D with stroke (đ U+0111)
- Latin letter E with acute (é U+00E9)
- Latin letter E with grave (è U+00E8)
- Latin letter E with circumflex (ê U+00EA)
- Latin letter Eng (ŋ U+014B)
- Latin letter N with acute (ń U+0144)
- Latin letter N with tilde (ñ U+00F1)
- Latin letter O with acute (ó U+00F3)
- Latin letter O with grave (ò U+00F2)
- Latin letter O with circumflex (ô U+00F4)
- Latin letter O with diaeresis (ö U+00F6)
- Latin letter S with caron (š U+0161)
- Latin letter T with stroke (ŧ U+0167)
- Latin letter U with diaeresis (ü U+00FC)
- Latin letter Z with caron (ž U+017E)
- Latin letter AE ligature (æ U+00E6)
- Latin letter O with stroke (ø U+00F8)
- Latin letter A with ring above (å U+00E5)

These accents have been further classified into regional groups. (*This will be documented in a future revision –kim*)

Source: NORID

Polish (PL)

- LDH
- Latin letter O with acute (ó U+00F3)
- Latin letter A with ogonek (ą U+0105)
- Latin letter C with acute (ć U+0107)
- Latin letter E with ogonek (ę U+0119)
- Latin letter L with stroke (ł U+0142)
- Latin letter N with acute (ń U+0144)
- Latin letter S with acute (ś U+015B)
- Latin letter Z with acute (ź U+017A)
- Latin letter Z with dot above (ż U+017C)

Source: NASK

Slovenian (SI)

- LDH
- Latin letter C with caron (č U+010D)
- Latin letter S with caron (š U+0161)
- Latin letter Z with caron (ž U+017E)

Source: ARNES

Spanish (ES)

The official language of Spain is Castilian (ES-ES), however other officially recognised local languages are Galician (ES-GL), Basque (ES-EU) and Catalan (ES-CA).

Code points required for Castilian are:

- LDH
- Latin letter A with acute ()
- Latin letter E with acute
- Latin letter I with acute
- Latin letter N with
- Latin letter O with acute
- Latin letter U with acute
- Latin letter U with diaeresis

Code points required for Catalan are:

- LDH
- Latin letter A with grave ()
- Latin letter C with cedilla
- Latin letter E with grave
- Latin letter I with diaeresis
- Latin letter O with grave

Source: Red.es

Swedish (SE)

- LDH
- Latin letter A with ring above (å U+00E5)
- Latin letter A with diaeresis (ä U+00E4)
- Latin letter E with acute (é U+00E9)
- Latin letter O with diaeresis (ö U+00F6)
- Latin letter O with diaeresis (ü U+00FC)

Source: NIC-SE

United Kingdom National Languages (UK)

The superset of code points aims to accommodate various national languages of the United Kingdom, including English, Welsh, Scottish Gaelic, Irish, Cornish/Kernewek, Scots/Lallans, and Ulster-Scots.

The modern transliterations of the minority languages principally use characters borrowed from the English alphabet with vowels (including ‘w’ and ‘y’) augmented by diacritical marks (diæresis, accent acute, accent grave and circumflex). Apostrophes are also significant.

- LDH
- Latin letter A with grave (à U+00E0)
- Latin letter A with acute (á U+00E1)
- Latin letter A with circumflex (â U+00E2)

- Latin letter A with diæresis (ä U+00E4)
- Latin letter E with grave (è U+00E8)
- Latin letter E with acute (é U+00E9)
- Latin letter E with circumflex (ê U+00EA)
- Latin letter E with diæresis (ë U+00EB)
- Latin letter I with grave (ì U+00EC)
- Latin letter I with acute (í U+00ED)
- Latin letter I with circumflex (î U+00EE)
- Latin letter I with diæresis (ï U+00EF)
- Latin letter O with grave (ò U+00F2)
- Latin letter O with acute (ó U+00F3)
- Latin letter O with circumflex (ô U+00F4)
- Latin letter O with diæresis (ö U+00F6)
- Latin letter U with grave (ù U+00F9)
- Latin letter U with acute (ú U+00FA)
- Latin letter U with circumflex (û U+00FB)
- Latin letter U with diæresis (ü U+00FC)
- Latin letter W with circumflex (ŵ U+0175)
- Latin letter Y with acute (ý U+00FD)
- Latin letter Y with circumflex (ŷ U+0177)
- Latin letter Y with diaeresis (ÿ U+00FF)
- Apostrophe (’ U+2019)

Source: Nominet UK

Combined Codepoint Matrix

Character		**Code Point**	**Languages**	**Pre-normalised Forms *(incomplete)***
Latin letter A with acute	á	U+00E1	Castilian Hungarian Icelandic Norwegian UK National Languages	U+00C1
Latin letter A with circumflex	â	U+00E2	French UK National Languages	U+00C2
Latin letter A with diaeresis	ä	U+00E4	German Danish Norwegian Swedish UK National Languages	U+00C4
Latin letter A with grave	à	U+00E0	Catalan French Norwegian UK National Languages	U+00C0
Latin letter A with ogonek	ą	U+0105	Lithuanian Polish	U+0104
Latin letter A with ring above	å	U+00E5	Danish Norwegian Swedish	U+00C5
Latin letter A	a	U+0061	LDH	U+0041
Latin letter AE ligature	æ	U+00E6	Danish French Icelandic Norwegian	U+00C6
Latin letter B	b	U+0062	LDH	U+0042
Latin letter C with acute	ć	U+0107	Polish	U+0106
Latin letter C with caron	č	U+010D	Lithuanian Norwegian Slovenian	U+010C
Latin letter C with cedilla	ç	U+00E7	Catalan French Norwegian	U+00C7
Latin letter C with dot above	ċ	U+010B	Maltese	U+010A
Latin letter C	c	U+0063	LDH	U+0043
Latin letter D with stroke	đ	U+0111	Norwegian	
Latin letter D	d	U+0064	LDH	U+0044

Latin letter E with acute	é	U+00E9	Danish Castilian French Hungarian Icelandic Norwegian Swedish UK National Languages	U+00C9
Latin letter E with circumflex	ê	U+00EA	French Norwegian UK National Languages	U+00CA
Latin letter E with diæresis	ë	U+00EB	French UK National Languages	U+00CB
Latin letter E with dot above	ė	U+0117	Lithuanian	U+0116
Latin letter E with grave	è	U+00E8	Catalan French Norwegian UK National Languages	U+00C8
Latin letter E with ogonek	ę	U+0119	Lithuanian Polish	U+0118
Latin letter E	e	U+0065	LDH	U+0045
Latin letter Eng	ŋ	U+014B	Norwegian	U+014A
Latin letter Eth	ð	U+00F0	Icelandic	U+00C0
Latin letter F	f	U+0066	LDH	U+0046
Latin letter G with dot above	ġ	U+0121	Maltese	U+0120
Latin letter G	g	U+0067	LDH	U+0047
Latin letter H with stroke	ħ	U+0127	Maltese	U+0126
Latin letter H	h	U+0068	LDH	U+0048
Latin letter I with acute	í	U+00ED	Castilian Hungarian Icelandic UK National Languages	U+00CD
Latin letter I with circumflex	î	U+00EE	French UK National Languages	U+00CE
Latin letter I with diaeresis	ï	U+00EF	Catalan French UK National Languages	U+00CF
Latin letter I with grave	ì	U+00EC	UK National Languages	U+00CC
Latin letter I with ogonek	į	U+012F	Lithuanian	U+012E
Latin letter I	i	U+0069	LDH	U+0049
Latin letter J	j	U+006a	LDH	U+004a

Latin letter K	k	U+006b	LDH	U+004b
Latin letter L with stroke	ł	U+0142	Polish	U+0141
Latin letter L	l	U+006c	LDH	U+004c
Latin letter M	m	U+006d	LDH	U+004d
Latin letter N with acute	ń	U+0144	Norwegian Polish	U+0143
Latin letter N with tilde	ñ	U+00F1	Castilian Norwegian	U+00C1 U+004E U+0303 U+006E U+0303
Latin letter N	n	U+006e	LDH	U+004e
Latin letter O with acute	ó	U+00F3	Castilian Hungarian Icelandic Norwegian Polish UK National Languages	U+00D3
Latin letter O with double acute	ő	U+0151	Hungarian	U+0150
Latin letter O with circumflex	ô	U+00F4	French Norwegian UK National Languages	U+00D4
Latin letter O with diaeresis	ö	U+00F6	German Danish Hungarian Icelandic Norwegian Swedish UK National Languages	U+00D6
Latin letter O with grave	ò	U+00F2	Catalan Norwegian UK National Languages	U+00D2
Latin letter O with stroke	ø	U+00F8	Danish Norwegian	U+00D8
Latin letter OE ligature	œ	U+0153	French	U+0152
Latin letter O	o	U+006f	LDH	U+004f
Latin letter P	p	U+0070	LDH	U+0050
Latin letter Q	q	U+0071	LDH	U+0051
Latin letter R	r	U+0072	LDH	U+0052
Latin letter S with acute	ś	U+015B	Polish	U+015A
Latin letter S with caron	š	U+0161	Lithuanian Norwegian Slovenian	U+0160
Latin letter S	s	U+0073	LDH	U+0053
Latin letter T with stroke	ŧ	U+0167	Norwegian	U+0166
Latin letter T	t	U+0074	LDH	U+0054

Latin letter U with acute	ú	U+00FA	Castilian Hungarian Icelandic UK National Languages	U+00DA
Latin letter U with double acute	ű	U+0171	Hungarian	U+0170
Latin letter U with circumflex	û	U+00FB	French UK National Languages	U+00DB
Latin letter U with diaeresis	ü	U+00FC	German Danish Castilian French Hungarian Norwegian UK National Languages	U+00DC
Latin letter U with grave	ù	U+00F9	French UK National Languages	U+00D9
Latin letter U with macron	ū	U+016B	Lithuanian	U+016A
Latin letter U with ogonek	ų	U+0173	Lithuanian	U+0172
Latin letter U	u	U+0075	LDH	U+0055
Latin letter V	v	U+0076	LDH	U+0056
Latin letter W with circumflex	ŵ	U+0175	UK National Languages	U+0174
Latin letter W	w	U+0077	LDH	U+0057
Latin letter X	x	U+0078	LDH	U+0058
Latin letter Y with acute	ý	U+00FD	Icelandic UK National Languages	U+00FC
Latin letter Y with circumflex	ŷ	U+0177	UK National Languages	U+0176
Latin letter Y with diæresis	ÿ	U+00FF	French UK National Languages	U+0178
Latin letter Y	y	U+0079	LDH	U+0059
Latin letter Z with acute	ź	U+017A	Polish	U+0179
Latin letter Z with caron	ž	U+017E	Lithuanian Norwegian Slovenian	U+017D
Latin letter Z with dot above	ż	U+017C	Maltese Polish	U+017B
Latin letter Z	z	U+007a	LDH	U+005a
Latin letter thorn	þ	U+00FE	Icelandic	U+00DE
Apostrophe	’	U+2019	UK National Languages	

Changes since version 0.2

- All the non code-points-per-language material has been stripped out. This should either be re-implemented in a different way, or, in a separate document dealing with actual deployment recommendations.

- Unicode references for Swedish codepoints added

- Information on Castilian and Catalan added, as provided by Red.es. General information on officially recognised Spanish languages also added.

- Reinserted French language

- Information about Hungarian added, provided by Hungarian registry.

- Information about Danish added, provided by Danish registry.