A Latin Character Repertoire for Internationalized Domain Names

Cary Karp
Swedish Museum of Natural History

1. Introduction

The registries for the .SE national top-level domain and the .MUSEUM generic top-level domain both have their administrative and operational headquarters in Stockholm, Sweden. The two registries are therefore collaborating on the implementation of Internationalized Domain Names ("IDNs") derived from all languages with official status in that country. (An IDN is a domain name containing at least one character that is not a basic Latin letter "a-z", a digit "0-9", or a hyphen "-", as defined in <http://www.faqs.org/rfcs/rfc3490.htm>. Readers unfamiliar with the basic concepts of domain naming will find a brief review in an appendix, below.)

In addition to Swedish, which is the de facto national language, there are five legally recognized minority languages: Finnish, Meänkeli, Romani, Sami, and Yiddish. Strict guidelines provided by the national government about names in administrative systems specify a Latin character repertoire that provides a comprehensive basis for the IDN representation of all of these languages, except Yiddish which is written with an alphabet based on Hebrew script.

The script table presented here was taken directly from the governmental guidelines at <http://about.museum/idn/riktlinjer0505.pdf>, which also include narrative text (in Swedish). The derivation of the corresponding language table for Yiddish IDNs in the .MUSEUM and .SE registries is described separately at <http://about.museum/idn/museum-se-yiddish.pdf>.

2. Policies

The table in the next section subsumes all of the language-specific character tables that have previously been placed in the *IANA Repository of TLD IDN Practices* by the .MUSEUM and .SE registries, and obsoletes all but one of them. The exception is the Swedish language table, which is being retained for the information of anyone interested in a listing of the IDN elements of that language, as deemed necessary by a registry that conducts its primary business in Swedish <http://www.iana.org/assignments/idn/se-swedish.html>.

The tabulated repertoire is an extensive listing of characters in the Unicode Character Code Chart <http://www.unicode.org/charts/> which are named there as belonging to the LATIN script, and combined with no more than one diacritical mark or similar graphic extension, with each composite character listed at a single numerical "code point". The repertoire supports numerous languages written using the Latin alphabet and is intended to permit the representation of names derived from European languages, using their native orthographies to the fullest extent possible. There is, however, neither a requirement nor an expectation that a label in a domain name will correspond to a proper name or dictionary word in any language, and many labels deliberately do not have any such attributes.

There is therefore no basis for determining the extent to which any word-based restrictions or other language-specific orthographic conventions can be applied here and, in consequence, all registration policies are script based. Any permissible character may appear at any point in a string, with the exception of digits and the hyphen, which may not be in the initial or final positions in a label.

The holder of an IDN is responsible for the orthographic rigor of any proper words or names used as labels. Each representation of a label in an alternative orthography requires separate registration. For example, the prospective holder of the label “lättöl” is free to register the correlate “lattol”, without either form imposing any restriction on the availability of the other, or on any further variants using the more than twenty diacritically marked forms of the base “a” in the Unicode chart, or the similar number of marked forms of “o”.

This also applies to marked or ligated characters that can alternately be represented as digraphs. It is again up to the prospective name holder to make an individual determination as to whether or not there is an equivalence between an umlauted “ä”, and an “ae” digraph or an “æ” ligature, or if the “ä” can acceptably also be indicated with an “a”.

Even if lexicographic rules might be contemplated for reducing the inherent ambiguity, their automated implementation would easily be stymied by reasonable differences between the representations of both proper names and dictionary words: “encyclopaedia” and “encyclopædia” could be treated as identical, but “mueller” and “müller” cannot, and “öresund” and “øresund” can be argued either way.

3. Character table

The value in the first column of this table gives the position of a character in the Unicode chart with “U+” prefixed to the code point in hexadecimal form. (The unprefixed “nnnn..mmm” in the supplementary table indicates a continuous range of code points.) The second column illustrates the corresponding characters. The third column gives the Unicode names for the characters. The code points are listed according to the European Ordering Rules (ENV 13710).

Code Point	Symbol	Unicode Name
U+0061	**a**	LATIN SMALL LETTER A
U+00E1	**á**	LATIN SMALL LETTER A WITH ACUTE
U+00E0	**à**	LATIN SMALL LETTER A WITH GRAVE
U+0103	**ă**	LATIN SMALL LETTER A WITH BREVE
U+00E2	**â**	LATIN SMALL LETTER A WITH CIRCUMFLEX
U+00E5	**å**	LATIN SMALL LETTER A WITH RING ABOVE
U+00E4	**ä**	LATIN SMALL LETTER A WITH DIAERESIS
U+00E3	**ã**	LATIN SMALL LETTER A WITH TILDE
U+0105	**ą**	LATIN SMALL LETTER A WITH OGONEK
U+0101	**ā**	LATIN SMALL LETTER A WITH MACRON

U+01CE	**ǎ**	LATIN SMALL LETTER A WITH CARON
U+00E6	**æ**	LATIN SMALL LETTER AE
U+0062	**b**	LATIN SMALL LETTER B
U+0063	**c**	LATIN SMALL LETTER C
U+0107	**ć**	LATIN SMALL LETTER C WITH ACUTE
U+010D	**č**	LATIN SMALL LETTER C WITH CARON
U+010B	**ċ**	LATIN SMALL LETTER C WITH DOT ABOVE
U+00E7	**ç**	LATIN SMALL LETTER C WITH CEDILLA
U+0064	**d**	LATIN SMALL LETTER D
U+010F	**ď**	LATIN SMALL LETTER D WITH CARON
U+0111	**đ**	LATIN SMALL LETTER D WITH STROKE
U+00F0	**ð**	LATIN SMALL LETTER ETH
U+0065	**e**	LATIN SMALL LETTER E
U+00E9	**é**	LATIN SMALL LETTER E WITH ACUTE
U+00E8	**è**	LATIN SMALL LETTER E WITH GRAVE
U+00EA	**ê**	LATIN SMALL LETTER E WITH CIRCUMFLEX
U+011B	**ě**	LATIN SMALL LETTER E WITH CARON
U+00EB	**ë**	LATIN SMALL LETTER E WITH DIAERESIS
U+0119	**ę**	LATIN SMALL LETTER E WITH OGONEK
U+0113	**ē**	LATIN SMALL LETTER E WITH MACRON
U+0117	**ė**	LATIN SMALL LETTER E WITH DOT ABOVE
U+0259	**ə**	LATIN SMALL LETTER SCHWA
U+0066	**f**	LATIN SMALL LETTER F
U+0067	**g**	LATIN SMALL LETTER G
U+011F	**ğ**	LATIN SMALL LETTER G WITH BREVE
U+01E7	**ǧ**	LATIN SMALL LETTER G WITH CARON
U+0121	**ġ**	LATIN SMALL LETTER G WITH DOT ABOVE
U+0123	**ģ**	LATIN SMALL LETTER G WITH CEDILLA
U+01E5	**ǥ**	LATIN SMALL LETTER G WITH STROKE
U+0068	**h**	LATIN SMALL LETTER H
U+0127	**ħ**	LATIN SMALL LETTER H WITH STROKE
U+0069	**i**	LATIN SMALL LETTER I
U+0131	**ı**	LATIN SMALL LETTER DOTLESS I
U+00ED	**í**	LATIN SMALL LETTER I WITH ACUTE
U+00EC	**ì**	LATIN SMALL LETTER I WITH GRAVE
U+00EE	**î**	LATIN SMALL LETTER I WITH CIRCUMFLEX
U+00EF	**ï**	LATIN SMALL LETTER I WITH DIAERESIS
U+012F	**į**	LATIN SMALL LETTER I WITH OGONEK
U+012B	**ī**	LATIN SMALL LETTER I WITH MACRON
U+01D0	**ǐ**	LATIN SMALL LETTER I WITH CARON
U+006A	**j**	LATIN SMALL LETTER J
U+006B	**k**	LATIN SMALL LETTER K
U+01E9	**ǩ**	LATIN SMALL LETTER K WITH CARON
U+0137	**ķ**	LATIN SMALL LETTER K WITH CEDILLA

U+006C	**l**	LATIN SMALL LETTER L
U+013A	**ĺ**	LATIN SMALL LETTER L WITH ACUTE
U+013E	**ľ**	LATIN SMALL LETTER L WITH CARON
U+013C	**ļ**	LATIN SMALL LETTER L WITH CEDILLA
U+0142	**ł**	LATIN SMALL LETTER L WITH STROKE
U+006D	**m**	LATIN SMALL LETTER M
U+006E	**n**	LATIN SMALL LETTER N
U+0144	**ń**	LATIN SMALL LETTER N WITH ACUTE
U+0148	**ň**	LATIN SMALL LETTER N WITH CARON
U+00F1	**ñ**	LATIN SMALL LETTER N WITH TILDE
U+0146	**ņ**	LATIN SMALL LETTER N WITH CEDILLA
U+014B	**ŋ**	LATIN SMALL LETTER ENG
U+006F	**o**	LATIN SMALL LETTER O
U+00F3	**ó**	LATIN SMALL LETTER O WITH ACUTE
U+00F2	**ò**	LATIN SMALL LETTER O WITH GRAVE
U+00F4	**ô**	LATIN SMALL LETTER O WITH CIRCUMFLEX
U+00F6	**ö**	LATIN SMALL LETTER O WITH DIAERESIS
U+0151	**ő**	LATIN SMALL LETTER O WITH DOUBLE ACUTE
U+00F5	**õ**	LATIN SMALL LETTER O WITH TILDE
U+014D	**ō**	LATIN SMALL LETTER O WITH MACRON
U+01D2	**ǒ**	LATIN SMALL LETTER O WITH CARON
U+00F8	**ø**	LATIN SMALL LETTER O WITH STROKE
U+0153	**œ**	LATIN SMALL LIGATURE OE
U+0070	**p**	LATIN SMALL LETTER P
U+0071	**q**	LATIN SMALL LETTER Q
U+0072	**r**	LATIN SMALL LETTER R
U+0155	**ŕ**	LATIN SMALL LETTER R WITH ACUTE
U+0159	**ř**	LATIN SMALL LETTER R WITH CARON
U+0157	**ŗ**	LATIN SMALL LETTER R WITH CEDILLA
U+0073	**s**	LATIN SMALL LETTER S
U+015B	**ś**	LATIN SMALL LETTER S WITH ACUTE
U+0161	**š**	LATIN SMALL LETTER S WITH CARON
U+015F	**ş**	LATIN SMALL LETTER S WITH CEDILLA
U+0074	**t**	LATIN SMALL LETTER T
U+0165	**ť**	LATIN SMALL LETTER T WITH CARON
U+0163	**ţ**	LATIN SMALL LETTER T WITH CEDILLA
U+0167	**ŧ**	LATIN SMALL LETTER T WITH STROKE
U+0075	**u**	LATIN SMALL LETTER U
U+00FA	**ú**	LATIN SMALL LETTER U WITH ACUTE
U+00F9	**ù**	LATIN SMALL LETTER U WITH GRAVE
U+00FB	**û**	LATIN SMALL LETTER U WITH CIRCUMFLEX
U+016F	**ů**	LATIN SMALL LETTER U WITH RING ABOVE
U+00FC	**ü**	LATIN SMALL LETTER U WITH DIAERESIS
U+0171	**ű**	LATIN SMALL LETTER U WITH DOUBLE ACUTE

U+0173	**ų**	LATIN SMALL LETTER U WITH OGONEK
U+016B	**ū**	LATIN SMALL LETTER U WITH MACRON
U+01D4	**ǔ**	LATIN SMALL LETTER U WITH CARON
U+0076	**v**	LATIN SMALL LETTER V
U+0077	**w**	LATIN SMALL LETTER W
U+1E83	**ẃ**	LATIN SMALL LETTER W WITH ACUTE
U+1E81	**ẁ**	LATIN SMALL LETTER W WITH GRAVE
U+0175	**ŵ**	LATIN SMALL LETTER W WITH CIRCUMFLEX
U+1E85	**ẅ**	LATIN SMALL LETTER W WITH DIAERESIS
U+0078	**x**	LATIN SMALL LETTER X
U+0079	**y**	LATIN SMALL LETTER Y
U+00FD	**ý**	LATIN SMALL LETTER Y WITH ACUTE
U+1EF3	**ỳ**	LATIN SMALL LETTER Y WITH GRAVE
U+0177	**ŷ**	LATIN SMALL LETTER Y WITH CIRCUMFLEX
U+00FF	**ÿ**	LATIN SMALL LETTER Y WITH DIAERESIS
U+007A	**z**	LATIN SMALL LETTER Z
U+017A	**ź**	LATIN SMALL LETTER Z WITH ACUTE
U+017E	**ž**	LATIN SMALL LETTER Z WITH CARON
U+017C	**ż**	LATIN SMALL LETTER Z WITH DOT ABOVE
U+0292	**ʒ**	LATIN SMALL LETTER EZH
U+01EF	**ǯ**	LATIN SMALL LETTER EZH WITH CARON
U+00FE	**þ**	LATIN SMALL LETTER THORN

A label containing a character at any of the code points specified above may not contain any other characters, except for those in the following auxiliary table, subject to constraints on the position in which they may appear:

U+002D	**-**	HYPHEN-MINUS
0030..0039	**0 - 9**	DIGIT ZERO – DIGIT NINE

Appendix Basic concepts

A domain name is the sequence of characters to the right of the @-sign in an e-mail address (username@**example.test**), or between the second and third slashes in a Web resource identifier (http://**example.test**/filename). It consists of “labels” separated by “dots”, with each label designating a level in the Domain Name System. In the second-level domain **example.test** (with both the labels and the dot being pronounced — “example dot test”) the top-level domain is **test**, (also commonly read with the preceding dot — “dot test”) and the second-level domain is **example**. This can be extended on successively lower levels as **fourthlevel.thirdlevel.secondlevel.toplevel**.

In e-mail addressing and Web resource identification, together with numerous other applications, the characters available for inclusion in a domain name are restricted to the twenty-six letters of the basic Latin alphabet “a-z”, the ten digits “0-9”, and the hyphen “-”. An Internationalized Domain Name appears to contain other characters, but this is done by encoding each IDN label with a sequence of characters taken from the restricted repertoire. For example, a label representing the Swedish word **lättöl** is normally displayed in the same form as an IDN, but it is actually stored in the Domain Name System in its encrypted form, which is **xn--lttl-loa4i**.

Software that understands this scheme displays the additional characters as a user expects to see them, transparently encoding and decoding them as required (sometimes needing explicit configuration before displaying the unencoded forms of characters that do not otherwise appear in the locale to which the software is set). A test site is maintained at <http://idn.icann.org/> to enable users to assess the IDN-compliance of their individual working environments in a number of different scripts and languages.

Version 1.0
12 January 2009

Author’s address:

Cary Karp
Swedish Museum of Natural History
Frescativägen 40
SE-11418 Stockholm