Considerations in the use of the Latin script in variant internationalized top-level domains

Final report of the ICANN VIP Study Group for the Latin script

Executive summary

The study group examined all the characters in the Unicode Character Code Chart version 6.1.0 that are associated with the Latin script and valid under the IDNA2008 protocol. It identified several forms of "confusability" that might require careful consideration in the collation of a subset of the broader repertoire for local use. The resolution of such issues is, however, highly dependent on local orthographic conventions. These frequently treat the same characters in different manners. Strings that are confusingly similar in the context of one language may have no such connotations in another.

Noting that the Latin script is used by a larger number of separate language communities than is any other single script, attempting to provide a comprehensive overview of the needs of all of them is an unrealistic endeavor. A summary attempt at doing so nonetheless would be culturally insensitive to communities that have yet to join the IDN discussion.

The study group therefore finds no basis for the categorical treatment of any code point assigned to an element of the Latin script as being equivalent to any other such code point. Nor does it believe that any such basis exists beyond what is already incorporated in the IDNA2008 protocol. The ICANN TLD application process should not permit requests for multiple Latin strings under the premise that they are variants of each other. Careful scrutiny is required when evaluating proposed TLD labels for confusability but that does not make them variants in the focused sense of the VIP study.

Two reference tables are appended to the report enumerating and illustrating the valid code points associated with the Latin script. The primary table lists those that can be taken as elements of everyday writing systems and are thus of potential utility in local IDN repertoires. The secondary table includes characters lacking that attribute or otherwise less suitable for unrestricted use. Although the former may be treated as the universe of code points available for prospective Latin-based TLD labels, the study group does not recommend the unconditional acceptance of any code point without specific demonstration of warrant for its use in the intended linguistic context. The Latin script table for the root zone can therefore be a descriptive collation of the code points actually appearing in that zone rather than a prescriptive tabulation of code points that can be included in it without further consideration.

1. Introduction

As part of its efforts to expand availability of internationalized domain names (IDNs), ICANN has initiated a limited program to introduce country-code top-level domains in non-Latin scripts[1] and is poised to launch a program that allows for generic top-level domains in any script.[2] One of the complexities of this process is that it challenges the assumption that a single domain can be represented with a single label in the Domain Name System. With IDNs, and the Unicode standard on which they are built, some labels can be represented in several forms. This occurs when an abstract character (see Section 6.1) in the label is found at more than one code point in the Unicode Character Code Chart or can be represented by alternative sequences of code points.

To facilitate the development of policies for the assignment and allocation of such multiple labels in the new top-level domain processes, ICANN has established the "IDN Variant TLD Project". Its mandate is to identify and analyze issues requiring consideration and provide recommendations upon which policies may than be based. As its first step, the study seeks to analyze comprehensively and come to a common understanding about what stakeholders consider to be variants and the relevant issues with them. These problem statements will be coalesced into a single requirements document against which solutions can be identified and assessed.

To obtain a broad cross-section of input, six case study teams were constituted — each for a particular script used in the writing systems of a prominent segment of the world's population. These scripts are Arabic, Chinese, Cyrillic, Devanagari, Greek, and Latin. The Latin case, which we analyze here, differs from the others in several respects. First, being the script on which the Domain Name System (DNS) is based, it has an almost 30-year history of deployment to support observations and conclusions. The inclusion of the basic "A to Z" Latin alphabet in the ASCII character set accommodates the needs of many language communities to a tolerable extent. The driving forces behind the use of the Latin script for IDNs can therefore differ from those for non-Latin scripts.

Second, as a script with strong historical ties to Cyrillic and Greek, commonalities must be considered among these three scripts. The practical deployment of multiple scripts entails the prior evaluation of confusability and security issues across their boundaries.

As a seed to their work, the Case Study teams were provided with a list of draft questions.[3] While these questions are not prescriptive, they were designed to touch on some of the key areas the teams were expected to consider in order to provide a comprehensive foundation for the problem statement.

In essence, the case study teams were asked to evaluate the concept of variants from the perspective of user experience. Users of the DNS — not just computer users, but also those using domains in mobile phones, system administrators maintaining technical systems, software vendors, registries, registrars, etc. — have expectations about how domain names should work. The introduction of variants should be based on filling defined gaps where those expectations need

---

1 http://www.icann.org/en/topics/idn/fast-track/

2 http://newgtlds.icann.org/

3 https://community.icann.org/download/attachments/16842778/Draft+Questions.pdf

to be met in order to have predictable and reliable operations. By outlining those expectation gaps as a problem statement, potential solutions can be evaluated in terms of their ability to address those gaps, balanced against their technical and policy implications.

2. Definitions

To ensure that the various study groups were using the same vocabulary, they were advised to adhere to specific terminology[4] wherever applicable. The definitions in RFC 6365[5] were to be considered as a base for internationalization terminology as enhancements to the provided definitions. Where that documentation does not clearly articulate a definition needed for this case study, further detail is added as appropriate in the present document.

3. The Latin script

The foundational repertoire of characters for the DNS is given in the ASCII standard. Although any octet may be used in a DNS name, a separate rule for host names limits the repertoire to the Latin letters "a" through "z" in both upper and lower case, the digits "0" through "9", and the hyphen "-" (the "LDH set"). Although there are differences in the way the DNS treats the upper- and lower case forms of the letters, they are largely regarded as transparent to users. These characters have been used for all registration of names in the DNS until the advent of IDNs in 2003.

The protocol used for implementing IDNs is formally known as "Internationalized Domain Names in Applications", or IDNA. This uses an "ASCII Compatible Encoding" for characters taken from the far broader Unicode repertoire. The encoding and decoding, as well as the display of Unicode characters is dependent on support on the application level. The corresponding registrations in the DNS, itself, remain ASCII based. The most recent version of this protocol is termed IDNA2008[6], (from the date of its initiation; it was published in 2010). Both it and the initial IDNA2003 rely solely on the Unicode standard[7] for deriving permissible code points.

The Unicode Standard assigns several character properties to each code point, of which one indicates script. This property is used to subdivide the full Unicode Character Chart into blocks headed by the script designator, as published on the Unicode website.[8]

The Latin script is presented in nine such blocks, listed in Appendix B1. The "Basic Latin" block restates the ASCII repertoire and therefore includes the LDH set to which host names were historically constrained. Latin letters beyond the LDH set, as well as diacritically-marked and

---

4 https://community.icann.org/download/attachments/16842778/Draft+Definitions.pdf

5 "Terminology Used in Internationalization in the IETF" (RFC 6365), http://tools.ietf.org/html/rfc6365

6 "Internationalised Domain Names in Applications (IDNA): Definitions and Document Framework" (RFC 5890) http://tools.ietf.org/html/rfc5890

7 http://unicode.org/standard/standard.html

8 http://www.unicode.org/charts/

otherwise decorated forms are presented in supplemental and extended Latin blocks, with further Latin letters in blocks under the heading “Phonetic Symbols”.

The IDNA2008 protocol permits the use of a large number of code points from these blocks (fully disallowing only the one headed “Fullwidth Latin Letters”). It also imposes some categorical limitations, for example, on code points that represent punctuation marks or symbols. The permissible code points are termed “protocol valid”, or PVALID.

The subset of protocol valid code points assigned to the Latin script which the group deemed relevant to the study of Latin variants is listed in Appendix B2. A second list of protocol valid code points that are not used in everyday writing systems but might nonetheless appear in IDNs under special circumstances is given in Appendix B3. Together they provide an exhaustive enumeration of what the group regarded as the Latin script code points.

4. User expectations

4.1. Case sensitivity

The Latin script basis for non-internationalized domain names has led to a well-established user expectation of how Latin script domain names function. In particular, the lack of distinction between upper- and lower case in LDH strings permits their artful representation in a manner that does not extend into IDNs. However, since users often type LDH labels in upper, lower, or mixed case it is reasonable to assume they will expect to be able to do so with other Latin characters, as well.

4.2. Web browsers

Web browsing software typically retains the case of an LDH domain name as it was entered, leaving the effects of any case normalization transparent to the user. This facility does not necessarily extend into the IDN environment and a number of other browser traits supporting IDNs lack counterpart in the traditional working environment. The difficulties users may have in accommodating themselves to this are compounded by the lack of a uniform approach to the implementation of IDNA by all browser developers. Some browsers display the ASCII encoded form of an IDN (the “A-label”) rather than the the expected localized form (the “U-label”) if the TLD is not “white listed” in a central repository maintained by the developer. Others browsers may expose the A-label based upon logic derived from the underlying operating system’s language settings, as well as those of the browser.

4.3. Keyboards

A-Z and 0-9 are present on most computer keyboards. Extended Latin characters that include diacritical marks, or are represented in other special forms, are represented on keyboards in regions where those characters are used. While there are minor differences in the layout of Latin keyboards, such as QWERTY or AZERTY, the LDH set is comprehensively available on all of them.

5. Relevant practices with non-IDN domains

Domain registries that allow IDN registration typically assemble smaller repertoires of permissible code points from the protocol-valid set. (In fact, IDNA2008 expects the collation of smaller local repertoires.) These are usually documented in tabular form, aggregated by script or language, as appropriate to the target community. Many are published in a central repository maintained by ICANN on the IANA website.[9]

There was no counterpart to this selectivity in prior practice. The LDH set provides twenty-six letters that can be supplemented with ten digits and a hyphen. The letters can be used without distinction between upper- and lower-case. The DNS protocol treats both forms as equivalent to the extent that "oldnews" and "OldNews" are the same label. This has allowed for visually similar labels such as "OldNews" versus "01dNews", as noted above. The prevalent use of fonts that obscure their visual differentiation can cause significant confusion (i.e. "OldNews" versus "0ldNews"). The group is unaware of attempts at limiting this by restricting permissible characters to an LDH subset, except for the root zone where numerical characters are excluded from domain names to avoid confusion with IP addresses.

Table: LDH Visual Ambiguities in the existing system[10]

G	6	
I	l	1
O	0	
S	5	
U	V	
Z	2	

5.2 Orthographic considerations

Differences between American and British spelling, such as "organization" and "organisation", are commonly noted but the group is unaware of their being included in any registry policies. The registration of a label in the one form confers no rights to the other, and a name holder wishing to obviate risk of user confusion has traditionally been free to register both. One of the reasons why such things were not perceived as a source of concern was that labels were regarded as mnemonic conveniences and any word-like properties they displayed were of secondary interest.

According to RFC 3454:

> "Users would probably expect all spelling equivalents to be made equivalent, or none of them to be. Examples of spelling equivalents include 'theater' vs. 'theatre', and 'hemoglobin' vs. 'hæmoglobin' in American vs. British English."[11]

[9] http://www.iana.org/domains/idn-tables

[10] Table sourced from "Punycode: A Bootstring encoding of Unicode for Internationalised Domain Names in Applications" (RFC 3492), Section 5. http://tools.ietf.org/html/rfc3492

[11] "Preparation of Internationalized Strings ('stringprep')" (RFC 3454). http://tools.ietf.org/html/rfc3454

> "Language-specific equivalences such as 'Aepfel' vs. 'Äpfel', which are sometimes considered equivalent in German, may not be considered equivalent in other languages."[12]

In the initial charter document for the Variant IDN Project, a Latin example was given of a hypothetical top-level domain for the city of Cologne (Köln), Germany. German spelling rules permit this to be written either as ".koeln" or ".köln".

For the purposes of the VIP study. the extent to which a German Internet user might view these two labels as similar, or to use another example, ".strasse" and ".straße", illustrates a situation that deserves consideration but is not amenable to algorithmic resolution. If anything, it highlights the extent to which such mathematical equivalence cannot be imposed on the Latin script.

6. Taxonomy of variants

Various terms have been used to designate the concept of equivalence but none has yet been provided with a definition that is adequate in all the contexts where it is needed. This terminological intractability is likely, at least in part, to be caused by differing perceptions and needs from one language community to another. This is true even among those language communities that share the same script. It is particularly true for the Latin script, which is used for writing a larger number of languages than all other scripts combined.

The most obvious variant relationship within the Latin alphabet is between the upper- and lower-case forms of a given letter. As noted, the DNS obviates need for concern with this, as long as the repertoire is restricted to the LDH set. The form of equality that it conveys to those characters is commonly taken to be somehow extensible into other variant relationships, recognizing that a correspondingly broader descriptive framework would be needed, but not suggesting any.

It is further commonly assumed that any greater complexity that might attach to the implementation of that speculative framework would not present an inordinate task to the technical community. In fact, ASCII case insensitivity is supported by a trivial mathematical operation. Devising means for the algorithmic enforcement of any other form of variant relationship — from the straightforward upper- and lower-case convertibility of Latin script letters beyond the LDH set, to the equivalence between characters that are considered similar in some other regard, is by no means a straightforward undertaking. If nothing else, characters used in the writing systems of multiple languages frequently have variant relationships that differ from one language to another, which would somehow also need universal quantification.

The IDNA protocol addresses central concerns with variant relationships in the Latin script, and deals, for example, directly with case sensitivity (albeit in what might be an unexpected manner). Other variant issues are so deeply dependent on local considerations that a general protocol-level solution is unlikely to be found.

---

[12] Ibid.

6.1. Display forms

An abstract character, say, the LATIN SMALL LETTER M (using the formal Unicode representation of character names), is instantiated by one or more glyphs, each with a specific display form. It is on this level that attributes such as serifs, italicizing, and weight are quantified:

m m *m* *m* **m** ***m***

A font is a named collection of glyphs (e.g. Helvetica or Times Roman), that will support one or more scripts and may use the same glyph in different scripts. A classic example of this is the LATIN SMALL LETTER A and the CYRILLIC SMALL LETTER A:

a а

Some fonts are deliberately designed to render such distinctions as visible as possible but the typical user is equipped with locale-specific fonts that do not normally support extensive subsets of the Unicode repertoire. The representation of different abstract characters with indistinguishable glyphs, using commonly available fonts at normal display resolutions, therefore weights heavily into other community-specific considerations when assessing the potential variant relationship between two code points.

6.2. Identical glyphs

The Latin script tables in Appendices B2 and B3 (explained in Section 6.4) contain only one instance of two abstract characters that typically share the same glyph. These are the LATIN SMALL LETTER TURNED E (U+01DD), and the LATIN SMALL LETTER SCHWA (U+0259):

ə ə

In any writing system in which either of them appears, the other is not included as a substitute. This allows for debate about the utility of indicating a variant relationship between them in any IDN repertoire in which they appear. The lack of locale-specific attributes in the root zone does, however, indicate a need for explicit contextual regulation of the use of these characters in TLD labels.

6.3. Case folding

The most immediate difference between the presentation of a string of Latin letters limited to the LDH set, and one in the wider IDNA-valid set, can be illustrated by returning to mixed-case labels. IDNA2008 does not operate on LDH labels and only accepts lower case letters in the labels that it does process. This means that the labels “clubamerica” and “ClubAmerica” match in the DNS, but “ClubAmérica” is invalid and can only be used as “clubamérica”.[13] In this case, the presence or absence of the accent is unlikely to change any meaning the label is intended to convey, and the name holder may ascribe greater value to flexibility in case representation than to the diacritical mark.

---

[13] IDNA2003 allows upper case input but normalizes it to lower case. IDNA2008 leaves software implementers free to include case-normalization in the preprocessing of IDNs.

There is less latitude in other situations involving no more than a single non-ASCII letter. Someone with the Swedish given name “Östen” and wishing to use it as a label in a domain name would almost certainly accept a restriction to lower case in preference to forgoing the umlaut and using the label “osten”, which is the Swedish word for cheese. In the German language, expanding an “ö” to “oe” is an established alternative, albeit of only limited applicability to proper names. However, this does not hold true for Swedish where the “ö” does not decompose into two letters, nor is considered a diacritically marked “o”. It is a letter in its own right (the final one in a 29-letter alphabet) and the two letters are not variants of one another. In text entry contexts, if an “ö” is unavailable the fallback is to use an “o”.

6.4. Decorative and contrastive variants

In cases where diacritical marks clarify pronunciation but do not change the meaning of a word — “naivete” versus “naïveté” — their use is “decorative”. In cases where the use of diacritical marks changes or disambiguates meaning — “resume” versus “résumé” — their use is “contrastive”. Irrespective of the distinction between decorative and contrastive usage, letters with diacritical marks are commonly referred to as “decorated”. The same term is also applied to letters that appear in modified shapes, whether or not they are marked:

ɲ ɳ ɲ

An extensive illustration of the decorated Latin letters available for use in IDNs, together with base letters that are not used in English and therefore not included in the LDH set, is given in the script tables in Appendices B2 and B3. These provide an exhaustive listing of the code points that are both labeled as letters in the Latin script blocks in Unicode version 6.1.0 (scheduled for release in February 2012 and not different in any relevant regard from version 6.0.0 currently in effect) and are valid under the IDNA2008 protocol. The phonetic blocks also include Latin elements that are used in writing systems which appear in the IDN space, or can reasonably be expected in the future. The tables currently include the code points from the phonetic blocks specifically needed to accommodate the “African Reference Alphabet”, or ARA. The ARA is used repeatedly in the illustrations here both because it is well suited for that purpose and to emphasize that concern with variant characters in the Latin script is not primarily anchored in European languages. Many of the code points included on its basis also appear in other Latin-based alphabets.

The Latin blocks were considered through the middle of Latin Extended-C, stopping at U+2C74. Code points beyond that are used in analytical discourse about language but do not appear in any everyday writing system of which the group is aware, with the single exception of U+A78C discussed below. These code points can be included in local repertoires, nonetheless, by registries that have documented need for them, understand how to embed them in strings that disambiguate their function (using label-based means for resolving character-level concerns) and are able to articulate the contextual constraints needed to prevent their inappropriate use elsewhere. It seems safe to assert that these code points currently have no utility warranting their inclusion in the root zone, and the group recommends that they be avoided entirely in that context.

6.5. Precomposed characters

The Unicode Latin blocks include a prodigious number of single code points that represent a base character combined with one or more diacritical marks, or appear in other decorated forms. Further marks can, however, be applied to these precomposed characters. No separate combining

marks are included in the table but many of the marks that appear in precomposed characters are also valid separately under the IDNA protocol in their combining forms. Anyone collating a local character table would therefore be free to include any of the following forms of the same character:

the precomposed,

LATIN SMALL LETTER C WITH CEDILLA AND ACUTE: ḉ

which can also be represented by combining characters at two separate code points,

LATIN SMALL LETTER C WITH CEDILLA + COMBINING ACUTE ACCENT: ḉ
LATIN SMALL LETTER C WITH ACUTE + COMBINING CEDILLA: ḉ

or at three separate code points,

LATIN SMALL LETTER C + COMBINING ACUTE ACCENT + COMBINING CEDILLA ḉ
LATIN SMALL LETTER C + COMBINING CEDILLA + COMBINING ACUTE ACCENT ḉ

These one-, two- and three code point representations are equivalent in every regard and thus an unambiguous instantiation of a variant character collection. The precomposed characters in the second and third of these examples can similarly be represented in both pre- and post-combined forms, providing two further collections. The number of variant character collections that can be culled recursively from the table in this manner would be truly overwhelming if such relationships were to be separately indicated. In practical terms this concern is largely, if not entirely, obviated by the normalization performed in IDNA2008.

6.6. Combining marks

As long as separate combining marks are excluded from the table there is no need to indicate the plethora of variant forms they can generate. There are, however, writing systems that use characters which the Unicode chart does not include in precomposed form. The ARA includes several underlined consonants of which some are available in precomposed form, such as the LATIN SMALL LETTER T WITH LINE BELOW (U+1E6F):

ṯ

with others requiring separate combining marks, such as the LATIN SMALL LETTER C + COMBINING LOW LINE (U+0063 U+0332):

c̲

The Unicode glossary refers to a composite character that does not exist in precomposed form as a “grapheme cluster”.[14]

[14] “A horizontally segmentable unit of text, consisting of some grapheme base ... together with any number of non-spacing markings applied to it.”

There are also situations where similar marks, such as the COMBINING LINE BELOW and the COMBINING MACRON BELOW, can be used interchangeably, with graphemically more distinct marks such the COMBINING DOT BELOW also being elements in what might be treated as a functional variant collection. There is, however, rarely any intrinsic variant relationship among the marks, themselves. That property is modulated by the base character with which they are combined, and further modulated by the writing system in which the combined characters are used. The treatment of marks as variants of each other in one writing system in no way precludes their being used independently of each other in another.

Combined characters can easily be placed in a table without requiring the separate listing of the combining marks, indicating variant relationships where appropriate. There must, however, be some form of contextual constraint placed on the character(s) to which a given mark may be applied. This is a good illustration of a situation where a massive problem in variant management can be averted on the protocol level. The initiation of such action is, however, not a present concern. The IDNA2008 protocol recognizes the need for reducing variant forms of this type to a single canonical representation and algorithmically normalizes all multi-code point alternatives to the precomposed form if one exists. Policy constraint is still necessary where it does not.

The five marked forms of the ‘c’ in the previous example, when included in a candidate U-label, appear indistinguishably as:

U+0061 U+1E09 U+0065 ; aḉe
U+0061 U+00E7 U+0301 U+0065 ; aḉe
U+0061 U+0107 U+0327 U+0065 ; aḉe
U+0061 U+0063 U+0301 U+0327 U+0065 ; aḉe
U+0061 U+0063 U+0327 U+0301 U+0065 ; aḉe

Although only the first of these is a valid U-label, the other forms are normalized to it before generating what is then the same A-label for them all — “xn--ae-ess”. It is important to note that this normalization is not performed by the encoding algorithm, itself. Each of the five candidate U-labels, valid and invalid, has its own Punycode form:

U+0061 U+1E09 U+0065 ; aḉe # ae-ess
U+0061 U+00E7 U+0301 U+0065 ; aḉe # ae-4ia66s
U+0061 U+0107 U+0327 U+0065 ; aḉe # ae-vla09s
U+0061 U+0063 U+0301 U+0327 U+0065 ; aḉe # ace-ldc1qd
U+0061 U+0063 U+0327 U+0301 U+0065 ; aḉe # ace-ldc3q

[Digression: There are online “IDN conversion” facilities that simply perform the Punycode encoding and prefix the result with ‘xn--’. It is therefore to be expected that some prospective name holders will submit requests with the A-label presented in an invalid form. As a further caveat, precomposed letters are often more stable typographically than are their multi-code point equivalents. Typical errors in the display of the latter are the misplacement of combining marks,

for example in the a◌̧e case, with either of the combining marks shifting from the ‘c’ to the ‘e’, without changing the actual code points. It is therefore to be expected that requests for U-labels may also be submitted in incorrect form. It is conceivable that the simultaneous appearance of both incorrect A- and U-labels will make it impossible to identify the actual label being requested.]

6.7. Punctuation

The single punctuation mark permitted in the LDH repertoire is the HYPHEN-MINUS (U+002D). The IDNA2008 repertoire includes a number of characters that are not punctuation marks but use glyphs that cannot be readily distinguished from them (and have Unicode names that are letter/punctuation hybrids). The use of such characters obviously requires extraordinary care. One example is the MODIFIER LETTER TURNED COMMA (U+02BB), which is used in numerous contexts of which one is as the “ ‘okina”, that indicates the glottal stop in, for example, “Hawai‘i”:

‘

Another example of a character used to indicate a glottal stop but which is a Latin letter and not a punctuation mark is the LATIN SMALL LETTER SALTILLO (U+A78C):

ꞌ

Its Unicode annotation states that, “it is widely used in many languages in Mexico and other regions, including Izere in Nigeria”, and it therefore may appear in localized Latin script tables. Similar expectations apply to the ‘okina in a Polynesian language context. The only basis for assembling these characters into a variant character collection is through their graphemic similarity. They are otherwise used in separate writing systems and are not mutually recognized across those boundaries. In any case, what would be the obvious preferred variant, the APOSTROPHE (U+0027), cannot appear in a U-label. The problem is that it is a shared fallback representation for the other forms, which are rarely indicated directly on stock keyboards and for which users commonly substitute an apostrophe. This all but guarantees a high failure rate when a label containing an apostrophe-like character is transcribed from ink on paper. The latitude for deceptive substitution in born-digital situations is similarly high.

The problem with characters that resemble apostrophes is compounded by the alternate rendering of the caron over some consonants either as the nominally expected mark seen in the first of the following illustrations or the apostrophe-like mark in the second. These two distinctly different renderings of the LATIN SMALL LETTER T WITH CARON (U+0165) provide a potential basis for yet another form of variant collection:

ť t’

An additional letter/punctuation similarity is found in the ARA. The LATIN LETTER GLOTTAL STOP (U+0294) resembles the QUESTION MARK (U+003F), which cannot appear in a U-label:

ʔ ?

Taking one of the four characters used to represent the click consonants in many languages spoken in Southern Africa (but not included in the ARA), the LATIN LETTER RETROFLEX CLICK

(U+01C3) similarly resembles, and in many fonts shares the same glyph with the EXCLAMATION MARK (U+0021), which is also barred from U-labels:

ǃ ǃ

For obvious reasons these characters, together with all other valid characters that resemble invalid punctuation marks, must not be permitted for use as surrogates for them.

6.8. Code points from unrelated writing systems

Similar concern is caused by letters that reside at different code points, are used in unrelated writing systems, and are commonly represented with similar, — if not identical — glyphs. Repeating the illustration of the LATIN SMALL LETTER TURNED E (U+01DD) and the LATIN SMALL LETTER SCHWA (U+0259) appearing above:

ǝ ə

The use of the same glyph for both in many widely deployed fonts does not imply that the characters are variant forms of each other, nor is any writing system likely to include more than one of them. Unforeseen problems might result, nonetheless, if a repertoire including one of these code points were to be compiled without awareness of the other. Therefore, in addition to any column in a generalized script table that indicates variant relationships between code points, a separate column could indicate code points that should not appear in the same localized repertoire without careful consideration. If both are indeed required, contextual rules for their use should be clearly articulated and enforced.

The converse also applies to characters within a single writing system. The ARA includes both the LATIN SMALL LETTER A (U+0061) and the LATIN SMALL LETTER ALPHA (U+0251), which cannot be taken as variants of each other:

a ɑ

However, the second of them clearly needs to be restricted to use in labels written with an alphabet that includes both. Here again, the inclusion of the latter in a TLD label will require explicit contextual regulation.

There is a daunting challenge in balancing need for minimizing the opportunity for abuse generated by the simultaneous availability of two characters such as these, against legitimate orthographic expectation — if not to say requirement — in linguistic contexts where the differences between the two letters are recognized and understood. This applies to many of the other illustrations given here.

The expertise needed to deal with such situations without any semblance of cultural discrimination extends beyond detailed first-hand knowledge of the respective language community’s writing system and keeping track of changes to its orthographic rules. It also requires in-depth familiarity with localized working environments, particularly in regard to the ability of commonly used rendering engines to display decorated characters correctly in widely deployed fonts, and awareness of any discrepancies between the code point for a given character as it is generated by a

stock keyboard and that character's Unicode code point. There is further need for close enough proximity to the community to be able to follow developments in the user environment and ensure the rapid adjustment of IDN systems to reflect changes as they are made.

If there is any variation within a language community in such things as the code point(s) produced for a given character on keyboard text entry, the preparation of a single IDN character repertoire for that community would also need to take those differences into account. (This is an important issue in variant management but not of particular concern with the Latin script.) Similar understanding of local conditions is necessary when assessing potential need for special treatment of characters at different code points which, at least in some contexts, are regularly displayed with glyphs that are indistinguishable to even the most erudite members of the speech communities using them.

Beyond familiarity with local conditions, assessing and managing these risks requires understanding the effects they can have on users external to a community who, on the one hand are comfortably able to read the basic script it uses for IDN labels, but on the other, will not recognize crucial code point distinctions projected onto familiar graphemes.

7. Visual similarity with other scripts

The group notes that some code points assigned to the Cyrillic and Greek scripts are commonly represented with glyphs that also are used for the display of code points assigned to the Latin script. Other scripts not included in the VIP study, such as Cherokee, have the same form of overlap with the Latin script.

One illustration of this property is the string “paypal”, written all in Latin script, versus “paypal”, using the Cyrillic “a”. These two strings are visually identical, not just similar, but are encoded differently.[15]

The Latin and Cyrillic Case Study Teams held a joint meeting to discuss the potential for confusion and harm with the mixed use of code points assigned to the two different scripts. The teams were not aware of any language written in either Latin or Cyrillic where the commingling of those scripts is required to express an idea or term, or otherwise has essential mnemonic value. Without a need for it, and given the risk for confusion that it engenders, the groups agreed that such commingling should be avoided. The thornier issue of “whole script confusability” — the juxtaposition of single-script labels that cannot be visually differentiated — was not discussed, as it was beyond the scope of the script-specific studies, and therefore expected to be considered in detail in the cross-study review.

8. Evaluation of TLD applications with variants

ICANN provides a tool for testing strings using the “SWORD algorithm”.[16] This has been online for preproduction testing during the period leading to the launch of the new gTLD program. Its

---

[15] “Draft Unicode Technical Standard 39: Unicode Security Mechanisms”, Section 4. http://unicode.org/reports/tr39/tr39-1.html

[16] http://icann.sword-group.com/icann-algorithm/Default.aspx

documentation claims support for a number of languages: "*This pre-production version algorithm supports the most common characters in Arabic, Chinese, Cyrillic, Devanagari, Greek, Japanese, Korean and Latin. ... Latin and Greek belong to the European script family and could be compared. ... This version of the algorithm does not include validation of candidate strings for compliance with IDNA protocols.*" [emphasis added]

There is an interest in having TLD applications vetted through this or other automated string evaluation mechanisms. Such a mechanism should flag and identify, say, two different applicants vying for .koeln and .köln, respectively. These might be considered a contention set, but suggesting this seemed beyond the charter or mandate of the initial drafting group.[17]

9. Impact of variants on registry/registrar operations

The management of TLD variants — i.e. the way a domain is delegated under a variant TLD — can raise issues at the registry and registrar levels. From a registry perspective, the following elements would have to be adjusted when introducing a variant: policies and procedures, including those applying to its launch and possible synchronization of domains under it; fees to the registrars; transfer procedures, including domain lock options.

At the registrar level, any development in the variant environment should foresee extended consultations with the accredited registrars of the registry that is planning to introduce a variant. There are multiple reasons for this of which the most pressing is because domains are registered via registrars and not directly with the registry in most TLDs. Launching a variant without any registrar involvement may lead to opportunities for benefit to a few privileged users. During the introduction of IDNs at the second level, many registrars did not immediately support scripts as they were introduced by a registry.

Therefore, the impact of variants in the registry/registrar field should be carefully explored not only at the level of operations — which should be investigated in any case — but most particularly at the point of their introduction, with primary focus on the interests of end users.

The impact of variant TLDs on registries and registrars may be highly dependent upon differing implementation methods. Any proposed implementation will require broad stakeholder participation to ensure that registries and registrars provide stable, secure, consistent, and unambiguous DNS operations. This includes the greatest possible clarity in communication and understanding of variant TLDs, to limit IDN end user, registrant, registrar, and registry confusion. Areas of application behavior, resolution and registration services, WHOIS service, and business logic all need to be examined in order to determine if these objectives are achievable.

10. Effects of variants on dispute resolution processes

10.1. Existing dispute resolution rules

Many TLD registries follow the Uniform Domain Name Dispute Resolution Policy (the "UDRP"). Under it, most types of trademark-based domain name disputes must be resolved by agreement,

---

[17] It should be noted that the SWORD site generates an error when these two strings are compared or köln is submitted for testing — *"köln : Label must be pure ASCII lowercase letters; found 'ö' (U+00F6) at position 1"*.

court action, or arbitration before a registrar will cancel, suspend, or transfer a domain name. Disputes alleged to arise from abusive registration of domain names (for example, cybersquatting) may be addressed by expedited administrative proceedings that the holder of trademark rights initiates by filing a complaint with an approved dispute resolution service provider.

An arbitration panel verifies whether the disputed domain (either the whole string, or a component of it) is identical or confusingly similar to a trademark or service mark in which the Complainant has rights, and whether it is creating a likelihood of confusion with the Complainant's mark as to the source, sponsorship, affiliation, or endorsement of the current registrant's website or location of a product or service on the current registrant's website or location.

Domains can be confusingly similar to the trademark or a service mark, taking into consideration:

> a) the part of the domain which is the label that was registered (typically the 2nd level part to the left of the top-level domain); or
> b) in rare cases, both the components to the left and the right of the dot (both the top-level and the second-level components of the domain).

10.2. Impact of Latin variants

In situations where variants for a single TLD are confusingly similar and the registry allows more than one party to register domains in the contention set, it is likely that the number of disputes will rise. For example, if a registry is granted two variant TLDs, say, ".koeln" and ".köln", and then delegates a domain under .koeln to one registrant, and the same label to another registrant under .köln — e.g. "hotel-airport.koeln" registered to one party, and "hotel-airport.köln" to another — there is a high likelihood that dispute resolution cases would arise from any policy that permitted this.

11. Conclusions

The conclusions of the Latin Case Study team are:

1. Several forms of visual similarity are found in the repertoire of Latin characters that can be used in IDNs. These require careful assessment when collating a subset of the full repertoire for local use. The resolution of the issues this can involve is, however, highly dependent on local orthographic conventions. These frequently treat the same characters in different manners. Strings that are confusingly similar in the context of one language may have no such connotations in another.

2. Since the Latin script is used by a larger number of separate language communities than is any other single script, attempting to provide a comprehensive overview of the needs of these communities is an unrealistic endeavor. A summary attempt at doing so nonetheless would be culturally insensitive to communities that have yet to join the IDN discussion.

3. Considering the two preceding points, the study group finds no basis for the categorical treatment of any code point assigned to an element of the Latin script as being equivalent to any other such code point. Nor does it believe that any such basis exists beyond what is already incorporated in the IDNA2008 protocol.

4. The ICANN TLD application process should not permit requests for multiple Latin strings under the premise that they are variants of each other. Careful scrutiny is required when evaluating proposed TLD labels for confusability but that does not make them variants in the focused sense of the VIP study.

5. Although the IDNA protocol permits a large number of code points to appear in Latin-based TLD labels, the study group does not recommend the unconditional acceptance of any code point for inclusion in the root zone of the DNS without specific demonstration of warrant in the intended linguistic context. In contrast to TLD zones that may have clear locale-specific attributes on which to assess such matters, the root is a globally shared resource and must accommodate all language communities in as equitable a manner as possible.

6. The Latin script table for the root zone can therefore be a descriptive collation of the code points actually appearing in that zone rather than a prescriptive tabulation of code points that can be included in it without further consideration.

APPENDICES

A. Latin Case Study Team

The Latin Case Study team was constituted, and first met, at the 41st ICANN meeting held in Singapore in June 2011. Subsequent to that meeting, it met roughly bi-weekly by telephone to consider the set of seed questions, discuss and analyze their relevance and implications regarding the Latin script, and to develop this report.

A1. Team membership

The Case Study team members are:

Name	Role
Jothan Frakes	Case Study Coordinator
Harald Alvestrand	Team Member
Andrzej Bartosiewicz	Team Member
Eric Brown	Team Member
Cary Karp	Team Member (representing the host organization)
Nadya Morozova	Team Member
Francisco Obispo	Team Member
Giovanni Seppia	Team Member
Will Shorter	Team Member
Wil Tan	Team Member
Avri Doria	Observer
Leo Vegoda	Case Study Liaison
Francisco Arias	Subject Matter Expert (Registry Operations)
Kim Davies	Subject Matter Expert (Security)
Nicholas Ostler	Subject Matter Expert (Linguistics)
Steve Sheng	Subject Matter Expert (Policy)
Andrew Sullivan	Subject Matter Export (Protocols)

A2. Declarations of interest

In order to ensure transparency by sharing relevant information on any interests the team's members have in relation to the areas of study, team members were asked to provide written statements declaring their interests.

These statements are published online at the case study team's website.[18]

A3. Acknowledgements

The team acknowledges the Internet Infrastructure Foundation, the host organization for the Latin Case Study.

18 https://community.icann.org/display/VIP/Latin

The team also acknowledges that key elements of this report were derived from an introductory Support Brief contributed by Cary Karp on behalf of the host organization, with the support of the Swedish Museum of Natural History.[19]

B. Latin character repertoire

In order to focus the work of the Case Study team — a definition of what characters are considered in scope needed to be made.

A limited subset of Unicode code points are considered valid IDNs according to the protocol specifications. These "protocol valid", or "PVALID", codes constrain which code points can be represented within the scope of the IDNA protocol.

This set can be further constrained to those code points that reflect those that belong to the various Unicode code point blocks that are ascribed to the Latin script.

B1. Unicode block assignment of code points to the Latin script

The Latin blocks in the Unicode Character Code Chart version 6.1.0 are:

Unicode Chart Name	Code point range
Basic Latin	U+0020 U+007F
Latin-1 Supplement	U+0080 U+00FF
Latin Extended-A	U+0100 U+017F
Latin Extended-B	U+0180 U+024F
Latin Extended-C	U+2C60 U+2C7F
Latin Extended-D	U+A720 U+A7FF
Latin Extended Additional	U+1E00 U+1EFF
Latin Ligatures	U+FB00 U+FB0F
Full-width Latin Letters	U+FF00 U+FF5E

B2. Protocol-valid code points assigned to the Latin script

The tables in this and the following section were collated according to the principles described in Section 6.4 in the main body of this report. These tables are intended to support and illustrate the action of the study and cannot be treated as authoritative in any regard. The normative statement of the protocol-valid code points is given in RFC 5892[20] with a corresponding reference table in the IANA Protocol Registry.[21]

Code points assigned to the Latin script that are expected to appear in locally collated IDN character repertoires follow in rough alphabetical order:

[19] http://mm.icann.org/pipermail/latin-vip/attachments/20111007/ed0ab5a4/latin-support-brief-final-0001.pdf

[20] http://tools.ietf.org/html/rfc5892

[21] http://www.iana.org/assignments/idnabis-tables/idnabis-tables.xml

U+0061 ; a # LATIN SMALL LETTER A
U+00E0 ; à # LATIN SMALL LETTER A WITH GRAVE
U+00E1 ; á # LATIN SMALL LETTER A WITH ACUTE
U+00E2 ; â # LATIN SMALL LETTER A WITH CIRCUMFLEX
U+00E3 ; ã # LATIN SMALL LETTER A WITH TILDE
U+00E4 ; ä # LATIN SMALL LETTER A WITH DIAERESIS
U+00E5 ; å # LATIN SMALL LETTER A WITH RING ABOVE
U+0101 ; ā # LATIN SMALL LETTER A WITH MACRON
U+0103 ; ă # LATIN SMALL LETTER A WITH BREVE
U+0105 ; ą # LATIN SMALL LETTER A WITH OGONEK
U+01CE ; ǎ # LATIN SMALL LETTER A WITH CARON
U+01DF ; ǟ # LATIN SMALL LETTER A WITH DIAERESIS AND MACRON
U+01E1 ; ǡ # LATIN SMALL LETTER A WITH DOT ABOVE AND MACRON
U+01FB ; ǻ # LATIN SMALL LETTER A WITH RING ABOVE AND ACUTE
U+0201 ; ȁ # LATIN SMALL LETTER A WITH DOUBLE GRAVE
U+0203 ; ȃ # LATIN SMALL LETTER A WITH INVERTED BREVE
U+1E01 ; ạ # LATIN SMALL LETTER A WITH RING BELOW
U+1EA1 ; ạ # LATIN SMALL LETTER A WITH DOT BELOW
U+1EA3 ; ả # LATIN SMALL LETTER A WITH HOOK ABOVE
U+1EA5 ; ấ # LATIN SMALL LETTER A WITH CIRCUMFLEX AND ACUTE
U+1EA7 ; ầ # LATIN SMALL LETTER A WITH CIRCUMFLEX AND GRAVE
U+1EA9 ; ẩ # LATIN SMALL LETTER A WITH CIRCUMFLEX AND HOOK
U+1EAB ; ẫ # LATIN SMALL LETTER A WITH CIRCUMFLEX AND TILDE
U+1EAD ; ậ # LATIN SMALL LETTER A WITH CIRCUMFLEX AND DOT BELOW
U+1EAF ; ắ # LATIN SMALL LETTER A WITH BREVE AND ACUTE
U+1EB1 ; ằ # LATIN SMALL LETTER A WITH BREVE AND GRAVE
U+1EB3 ; ẳ # LATIN SMALL LETTER A WITH BREVE AND HOOK ABOVE
U+1EB5 ; ẵ # LATIN SMALL LETTER A WITH BREVE AND TILDE
U+1EB7 ; ặ # LATIN SMALL LETTER A WITH BREVE AND DOT BELOW
U+0227 ; ȧ # LATIN SMALL LETTER A WITH DOT ABOVE
U+2C65 ; ⱥ # LATIN SMALL LETTER A WITH STROKE
U+0251 ; ɑ # LATIN SMALL LETTER ALPHA
U+00E6 ; æ # LATIN SMALL LETTER AE
U+0062 ; b # LATIN SMALL LETTER B

U+0180 ; ƀ # LATIN SMALL LETTER B WITH STROKE
U+0183 ; ƃ # LATIN SMALL LETTER B WITH TOPBAR
U+1E03 ; ḃ # LATIN SMALL LETTER B WITH DOT ABOVE
U+1E05 ; ḅ # LATIN SMALL LETTER B WITH DOT BELOW
U+1E07 ; ḇ # LATIN SMALL LETTER B WITH LINE BELOW
U+0253 ; ɓ # LATIN SMALL LETTER B WITH HOOK
U+0063 ; c # LATIN SMALL LETTER C
U+00E7 ; ç # LATIN SMALL LETTER C WITH CEDILLA
U+0107 ; ć # LATIN SMALL LETTER C WITH ACUTE
U+0109 ; ĉ # LATIN SMALL LETTER C WITH CIRCUMFLEX
U+010B ; ċ # LATIN SMALL LETTER C WITH DOT ABOVE
U+010D ; č # LATIN SMALL LETTER C WITH CARON
U+0188 ; ƈ # LATIN SMALL LETTER C WITH HOOK
U+023C ; ȼ # LATIN SMALL LETTER C WITH STROKE
U+0255 ; ɕ # LATIN SMALL LETTER C WITH CURL
U+A793 ; ꞓ # LATIN SMALL LETTER C WITH BAR
U+1E09 ; ḉ # LATIN SMALL LETTER C WITH CEDILLA AND ACUTE
U+2184 ; ↄ # LATIN SMALL LETTER REVERSED C
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+018C ; ƌ # LATIN SMALL LETTER D WITH TOPBAR
U+0221 ; ȡ # LATIN SMALL LETTER D WITH CURL
U+1E0D ; ḍ # LATIN SMALL LETTER D WITH DOT BELOW
U+1E0F ; ḏ # LATIN SMALL LETTER D WITH LINE BELOW
U+1E11 ; ḑ # LATIN SMALL LETTER D WITH CEDILLA
U+1E13 ; ḓ # LATIN SMALL LETTER D WITH CIRCUMFLEX BELOW
U+1E0B ; ḋ # LATIN SMALL LETTER D WITH DOT ABOVE
U+0256 ; ɖ # LATIN SMALL LETTER D WITH TAIL
U+0257 ; ɗ # LATIN SMALL LETTER D WITH HOOK
U+1E9F ; ẟ # LATIN SMALL LETTER DELTA
U+00F0 ; ð # LATIN SMALL LETTER ETH
U+0065 ; e # LATIN SMALL LETTER E
U+00E8 ; è # LATIN SMALL LETTER E WITH GRAVE

U+00E9 ; é # LATIN SMALL LETTER E WITH ACUTE
U+00EA ; ê # LATIN SMALL LETTER E WITH CIRCUMFLEX
U+00EB ; ë # LATIN SMALL LETTER E WITH DIAERESIS
U+0113 ; ē # LATIN SMALL LETTER E WITH MACRON
U+0115 ; ĕ # LATIN SMALL LETTER E WITH BREVE
U+0117 ; ė # LATIN SMALL LETTER E WITH DOT ABOVE
U+0119 ; ę # LATIN SMALL LETTER E WITH OGONEK
U+011B ; ě # LATIN SMALL LETTER E WITH CARON
U+0205 ; ȅ # LATIN SMALL LETTER E WITH DOUBLE GRAVE
U+0207 ; ȇ # LATIN SMALL LETTER E WITH INVERTED BREVE
U+1E15 ; ḕ # LATIN SMALL LETTER E WITH MACRON AND GRAVE
U+1E17 ; ḗ # LATIN SMALL LETTER E WITH MACRON AND ACUTE
U+1E19 ; ḙ # LATIN SMALL LETTER E WITH CIRCUMFLEX BELOW
U+1E1B ; ḛ # LATIN SMALL LETTER E WITH TILDE BELOW
U+1E1D ; ḝ # LATIN SMALL LETTER E WITH CEDILLA AND BREVE
U+1EB9 ; ẹ # LATIN SMALL LETTER E WITH DOT BELOW
U+1EBB ; ẻ # LATIN SMALL LETTER E WITH HOOK ABOVE
U+1EBD ; ẽ # LATIN SMALL LETTER E WITH TILDE
U+1EBF ; ế # LATIN SMALL LETTER E WITH CIRCUMFLEX AND ACUTE
U+1EC1 ; ề # LATIN SMALL LETTER E WITH CIRCUMFLEX AND GRAVE
U+1EC3 ; ể # LATIN SMALL LETTER E WITH CIRCUMFLEX AND HOOK ABOVE
U+1EC5 ; ễ # LATIN SMALL LETTER E WITH CIRCUMFLEX AND TILDE
U+1EC7 ; ệ # LATIN SMALL LETTER E WITH CIRCUMFLEX AND DOT BELOW
U+0229 ; ȩ # LATIN SMALL LETTER E WITH CEDILLA
U+0247 ; ɇ # LATIN SMALL LETTER E WITH STROKE
U+025B ; ɛ # LATIN SMALL LETTER OPEN E
U+01DD ; ǝ # LATIN SMALL LETTER TURNED E
U+0259 ; ə # LATIN SMALL LETTER SCHWA
U+0066 ; f # LATIN SMALL LETTER F
U+0192 ; ƒ # LATIN SMALL LETTER F WITH HOOK
U+1E1F ; ḟ # LATIN SMALL LETTER F WITH DOT ABOVE
U+0067 ; g # LATIN SMALL LETTER G
U+011D ; ĝ # LATIN SMALL LETTER G WITH CIRCUMFLEX
U+011F ; ğ # LATIN SMALL LETTER G WITH BREVE

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+01E7 ; ǧ # LATIN SMALL LETTER G WITH CARON
U+01F5 ; ǵ # LATIN SMALL LETTER G WITH ACUTE
U+0260 ; ɠ # LATIN SMALL LETTER G WITH HOOK
U+1E21 ; ḡ # LATIN SMALL LETTER G WITH MACRON
U+1D77 ; ᵷ # LATIN SMALL LETTER TURNED G
U+0263 ; ɣ # LATIN SMALL LETTER GAMMA
U+0068 ; h # LATIN SMALL LETTER H
U+0125 ; ĥ # LATIN SMALL LETTER H WITH CIRCUMFLEX
U+0127 ; ħ # LATIN SMALL LETTER H WITH STROKE
U+021F ; ȟ # LATIN SMALL LETTER H WITH CARON
U+1E23 ; ḣ # LATIN SMALL LETTER H WITH DOT ABOVE
U+1E25 ; ḥ # LATIN SMALL LETTER H WITH DOT BELOW
U+1E27 ; ḧ # LATIN SMALL LETTER H WITH DIAERESIS
U+1E29 ; ḩ # LATIN SMALL LETTER H WITH CEDILLA
U+1E2B ; ḫ # LATIN SMALL LETTER H WITH BREVE BELOW
U+2C68 ; ⱨ # LATIN SMALL LETTER H WITH DESCENDER
U+1E96 ; ẖ # LATIN SMALL LETTER H WITH LINE BELOW
U+0266 ; ɦ # LATIN SMALL LETTER H WITH HOOK
U+02AE ; ʮ # LATIN SMALL LETTER TURNED H WITH FISHHOOK
U+02AF ; ʯ # LATIN SMALL LETTER TURNED H WITH FISHHOOK AND TAIL
U+0069 ; i # LATIN SMALL LETTER I
U+00EC ; ì # LATIN SMALL LETTER I WITH GRAVE
U+00ED ; í # LATIN SMALL LETTER I WITH ACUTE
U+00EE ; î # LATIN SMALL LETTER I WITH CIRCUMFLEX
U+00EF ; ï # LATIN SMALL LETTER I WITH DIAERESIS
U+0129 ; ĩ # LATIN SMALL LETTER I WITH TILDE
U+012B ; ī # LATIN SMALL LETTER I WITH MACRON
U+012D ; ĭ # LATIN SMALL LETTER I WITH BREVE
U+012F ; į # LATIN SMALL LETTER I WITH OGONEK
U+0131 ; ı # LATIN SMALL LETTER DOTLESS I
U+01D0 ; ǐ # LATIN SMALL LETTER I WITH CARON

U+0209 ; ȉ # LATIN SMALL LETTER I WITH DOUBLE GRAVE
U+020B ; ȋ # LATIN SMALL LETTER I WITH INVERTED BREVE
U+1E2D ; ḭ # LATIN SMALL LETTER I WITH TILDE BELOW
U+1E2F ; ḯ # LATIN SMALL LETTER I WITH DIAERESIS AND ACUTE
U+1EC9 ; ỉ # LATIN SMALL LETTER I WITH HOOK ABOVE
U+1ECB ; ị # LATIN SMALL LETTER I WITH DOT BELOW
U+0269 ; ɩ # LATIN SMALL LETTER IOTA
U+006A ; j # LATIN SMALL LETTER J
U+0135 ; ĵ # LATIN SMALL LETTER J WITH CIRCUMFLEX
U+01F0 ; ǰ # LATIN SMALL LETTER J WITH CARON
U+0237 ; ȷ # LATIN SMALL LETTER DOTLESS J
U+0249 ; ɉ # LATIN SMALL LETTER J WITH STROKE
U+025F ; ɟ # LATIN SMALL LETTER DOTLESS J WITH STROKE
U+006B ; k # LATIN SMALL LETTER K
U+0137 ; ķ # LATIN SMALL LETTER K WITH CEDILLA
U+0199 ; ƙ # LATIN SMALL LETTER K WITH HOOK
U+01E9 ; ǩ # LATIN SMALL LETTER K WITH CARON
U+1E31 ; ḱ # LATIN SMALL LETTER K WITH ACUTE
U+1E33 ; ḳ # LATIN SMALL LETTER K WITH DOT BELOW
U+1E35 ; ḵ # LATIN SMALL LETTER K WITH LINE BELOW
U+2C6A ; ⱪ # LATIN SMALL LETTER K WITH DESCENDER
U+0138 ; ĸ # LATIN SMALL LETTER KRA
U+006C ; l # LATIN SMALL LETTER L
U+013A ; ĺ # LATIN SMALL LETTER L WITH ACUTE
U+013C ; ļ # LATIN SMALL LETTER L WITH CEDILLA
U+013E ; ľ # LATIN SMALL LETTER L WITH CARON
U+0142 ; ł # LATIN SMALL LETTER L WITH STROKE
U+019A ; ƚ # LATIN SMALL LETTER L WITH BAR
U+0234 ; ȴ # LATIN SMALL LETTER L WITH CURL
U+1E37 ; ḷ # LATIN SMALL LETTER L WITH DOT BELOW
U+1E39 ; ḹ # LATIN SMALL LETTER L WITH DOT BELOW AND MACRON
U+1E3B ; ḻ # LATIN SMALL LETTER L WITH LINE BELOW
U+1E3D ; ḽ # LATIN SMALL LETTER L WITH CIRCUMFLEX BELOW
U+2C61 ; ⱡ # LATIN SMALL LETTER L WITH DOUBLE BAR

U+006D ; m # LATIN SMALL LETTER M
U+1E3F ; ḿ # LATIN SMALL LETTER M WITH ACUTE
U+1E41 ; ṁ # LATIN SMALL LETTER M WITH DOT ABOVE
U+1E43 ; ṃ # LATIN SMALL LETTER M WITH DOT BELOW
U+006E ; n # LATIN SMALL LETTER N
U+00F1 ; ñ # LATIN SMALL LETTER N WITH TILDE
U+0144 ; ń # LATIN SMALL LETTER N WITH ACUTE
U+0146 ; ņ # LATIN SMALL LETTER N WITH CEDILLA
U+0148 ; ň # LATIN SMALL LETTER N WITH CARON
U+019E ; ƞ # LATIN SMALL LETTER N WITH LONG RIGHT LEG
U+01F9 ; ǹ # LATIN SMALL LETTER N WITH GRAVE
U+0235 ; n̯ # LATIN SMALL LETTER N WITH CURL
U+1E45 ; ṅ # LATIN SMALL LETTER N WITH DOT ABOVE
U+1E47 ; ṇ # LATIN SMALL LETTER N WITH DOT BELOW
U+1E49 ; n̲ # LATIN SMALL LETTER N WITH LINE BELOW
U+1E4B ; n̰ # LATIN SMALL LETTER N WITH CIRCUMFLEX BELOW
U+0272 ; ɲ # LATIN SMALL LETTER N WITH LEFT HOOK
U+014B ; ŋ # LATIN SMALL LETTER ENG
U+006F ; o # LATIN SMALL LETTER O
U+00F2 ; ò # LATIN SMALL LETTER O WITH GRAVE
U+00F3 ; ó # LATIN SMALL LETTER O WITH ACUTE
U+00F4 ; ô # LATIN SMALL LETTER O WITH CIRCUMFLEX
U+00F5 ; õ # LATIN SMALL LETTER O WITH TILDE
U+00F6 ; ö # LATIN SMALL LETTER O WITH DIAERESIS
U+00F8 ; ø # LATIN SMALL LETTER O WITH STROKE
U+014D ; ō # LATIN SMALL LETTER O WITH MACRON
U+014F ; ŏ # LATIN SMALL LETTER O WITH BREVE
U+0151 ; ő # LATIN SMALL LETTER O WITH DOUBLE ACUTE
U+01A1 ; ơ # LATIN SMALL LETTER O WITH HORN
U+01D2 ; ǒ # LATIN SMALL LETTER O WITH CARON
U+01EB ; ǫ # LATIN SMALL LETTER O WITH OGONEK
U+01ED ; ǭ # LATIN SMALL LETTER O WITH OGONEK AND MACRON
U+01FF ; ǿ # LATIN SMALL LETTER O WITH STROKE AND ACUTE
U+020D ; ȍ # LATIN SMALL LETTER O WITH DOUBLE GRAVE

U+020F ; ô # LATIN SMALL LETTER O WITH INVERTED BREVE
U+022B ; ȫ # LATIN SMALL LETTER O WITH DIAERESIS AND MACRON
U+022D ; ȭ # LATIN SMALL LETTER O WITH TILDE AND MACRON
U+022F ; ȯ # LATIN SMALL LETTER O WITH DOT ABOVE
U+0231 ; ȱ # LATIN SMALL LETTER O WITH DOT ABOVE AND MACRON
U+1E4D ; ố # LATIN SMALL LETTER O WITH TILDE AND ACUTE
U+1E4F ; ỗ # LATIN SMALL LETTER O WITH TILDE AND DIAERESIS
U+1E51 ; ồ # LATIN SMALL LETTER O WITH MACRON AND GRAVE
U+1E53 ; ố # LATIN SMALL LETTER O WITH MACRON AND ACUTE
U+1ECD ; ọ # LATIN SMALL LETTER O WITH DOT BELOW
U+1ECF ; ỏ # LATIN SMALL LETTER O WITH HOOK ABOVE
U+1ED1 ; ố # LATIN SMALL LETTER O WITH CIRCUMFLEX AND ACUTE
U+1ED3 ; ồ # LATIN SMALL LETTER O WITH CIRCUMFLEX AND GRAVE
U+1ED5 ; ổ # LATIN SMALL LETTER O WITH CIRCUMFLEX AND HOOK ABOVE
U+1ED7 ; ỗ # LATIN SMALL LETTER O WITH CIRCUMFLEX AND TILDE
U+1ED9 ; ộ # LATIN SMALL LETTER O WITH CIRCUMFLEX AND DOT BELOW
U+1EDB ; ớ # LATIN SMALL LETTER O WITH HORN AND ACUTE
U+1EDD ; ờ # LATIN SMALL LETTER O WITH HORN AND GRAVE
U+1EDF ; ở # LATIN SMALL LETTER O WITH HORN AND HOOK ABOVE
U+1EE1 ; ỡ # LATIN SMALL LETTER O WITH HORN AND TILDE
U+1EE3 ; ợ # LATIN SMALL LETTER O WITH HORN AND DOT BELOW
U+0254 ; ɔ # LATIN SMALL LETTER OPEN O
U+0275 ; ɵ # LATIN SMALL LETTER BARRED O
U+0153 ; œ # LATIN SMALL LIGATURE OE
U+0070 ; p # LATIN SMALL LETTER P
U+01A5 ; ƥ # LATIN SMALL LETTER P WITH HOOK
U+1E55 ; ṕ # LATIN SMALL LETTER P WITH ACUTE
U+1E57 ; ṗ # LATIN SMALL LETTER P WITH DOT ABOVE
U+0071 ; q # LATIN SMALL LETTER Q
U+024B ; ɋ # LATIN SMALL LETTER Q WITH HOOK TAIL
U+0072 ; r # LATIN SMALL LETTER R
U+0155 ; ŕ # LATIN SMALL LETTER R WITH ACUTE
U+0157 ; ŗ # LATIN SMALL LETTER R WITH CEDILLA
U+0159 ; ř # LATIN SMALL LETTER R WITH CARON

U+0211 ; ȑ # LATIN SMALL LETTER R WITH DOUBLE GRAVE
U+0213 ; ȓ # LATIN SMALL LETTER R WITH INVERTED BREVE
U+1E59 ; ṙ # LATIN SMALL LETTER R WITH DOT ABOVE
U+1E5B ; ṛ # LATIN SMALL LETTER R WITH DOT BELOW
U+1E5D ; ṝ # LATIN SMALL LETTER R WITH DOT BELOW AND MACRON
U+1E5F ; r̲ # LATIN SMALL LETTER R WITH LINE BELOW
U+024D ; ɍ # LATIN SMALL LETTER R WITH STROKE
U+027D ; ɽ # LATIN SMALL LETTER R WITH TAIL
U+027F ; ɿ # LATIN SMALL LETTER REVERSED R WITH FISHHOOK
U+0073 ; s # LATIN SMALL LETTER S
U+015B ; ś # LATIN SMALL LETTER S WITH ACUTE
U+015D ; ŝ # LATIN SMALL LETTER S WITH CIRCUMFLEX
U+015F ; ş # LATIN SMALL LETTER S WITH CEDILLA
U+0161 ; š # LATIN SMALL LETTER S WITH CARON
U+0219 ; ș # LATIN SMALL LETTER S WITH COMMA BELOW
U+1E61 ; ṡ # LATIN SMALL LETTER S WITH DOT ABOVE
U+1E63 ; ṣ # LATIN SMALL LETTER S WITH DOT BELOW
U+1E65 ; ṥ # LATIN SMALL LETTER S WITH ACUTE AND DOT ABOVE
U+1E67 ; ṧ # LATIN SMALL LETTER S WITH CARON AND DOT ABOVE
U+1E69 ; ṩ # LATIN SMALL LETTER S WITH DOT BELOW AND DOT
U+023F ; ȿ # LATIN SMALL LETTER S WITH SWASH TAIL
U+00DF ; ß # LATIN SMALL LETTER SHARP S
U+1E9C ; ẜ # LATIN SMALL LETTER LONG S WITH DIAGONAL STROKE
U+1E9D ; ẝ # LATIN SMALL LETTER LONG S WITH HIGH STROKE
U+0283 ; ʃ # LATIN SMALL LETTER ESH
U+0074 ; t # LATIN SMALL LETTER T
U+0163 ; ţ # LATIN SMALL LETTER T WITH CEDILLA
U+0165 ; ť # LATIN SMALL LETTER T WITH CARON
U+0167 ; ŧ # LATIN SMALL LETTER T WITH STROKE
U+01AB ; ƫ # LATIN LETTER T WITH PALATAL HOOK
U+01AD ; ƭ # LATIN SMALL LETTER T WITH HOOK
U+1E6B ; ṫ # LATIN SMALL LETTER T WITH DOT ABOVE
U+1E6D ; ṭ # LATIN SMALL LETTER T WITH DOT BELOW
U+1E6F ; t̲ # LATIN SMALL LETTER T WITH LINE BELOW

U+1E71 ; ṱ # LATIN SMALL LETTER T WITH CIRCUMFLEX BELOW
U+0236 ; ȶ # LATIN SMALL LETTER T WITH CURL
U+021B ; ț # LATIN SMALL LETTER T WITH COMMA BELOW
U+1E97 ; ẗ # LATIN SMALL LETTER T WITH DIAERESIS
U+2C66 ; ⱦ # LATIN SMALL LETTER T WITH DIAGONAL STROKE
U+0288 ; ʈ # LATIN SMALL LETTER T WITH RETROFLEX HOOK
U+00FE ; þ # LATIN SMALL LETTER THORN
U+0075 ; u # LATIN SMALL LETTER U
U+00F9 ; ù # LATIN SMALL LETTER U WITH GRAVE
U+00FA ; ú # LATIN SMALL LETTER U WITH ACUTE
U+00FB ; û # LATIN SMALL LETTER U WITH CIRCUMFLEX
U+00FC ; ü # LATIN SMALL LETTER U WITH DIAERESIS
U+0169 ; ũ # LATIN SMALL LETTER U WITH TILDE
U+016B ; ū # LATIN SMALL LETTER U WITH MACRON
U+016D ; ŭ # LATIN SMALL LETTER U WITH BREVE
U+016F ; ů # LATIN SMALL LETTER U WITH RING ABOVE
U+0171 ; ű # LATIN SMALL LETTER U WITH DOUBLE ACUTE
U+0173 ; ų # LATIN SMALL LETTER U WITH OGONEK
U+01B0 ; ư # LATIN SMALL LETTER U WITH HORN
U+01D4 ; ǔ # LATIN SMALL LETTER U WITH CARON
U+01D6 ; ǖ̄ # LATIN SMALL LETTER U WITH DIAERESIS AND MACRON
U+01D8 ; ǘ́ # LATIN SMALL LETTER U WITH DIAERESIS AND ACUTE
U+01DA ; ǚ̌ # LATIN SMALL LETTER U WITH DIAERESIS AND CARON
U+01DC ; ǜ̀ # LATIN SMALL LETTER U WITH DIAERESIS AND GRAVE
U+0215 ; ȕ # LATIN SMALL LETTER U WITH DOUBLE GRAVE
U+0217 ; û # LATIN SMALL LETTER U WITH INVERTED BREVE
U+1E73 ; ṳ # LATIN SMALL LETTER U WITH DIAERESIS BELOW
U+1E75 ; ṵ # LATIN SMALL LETTER U WITH TILDE BELOW
U+1E77 ; ṷ # LATIN SMALL LETTER U WITH CIRCUMFLEX BELOW
U+1E79 ; ṹ # LATIN SMALL LETTER U WITH TILDE AND ACUTE
U+1E7B ; ṻ # LATIN SMALL LETTER U WITH MACRON AND DIAERESIS
U+1EE5 ; ụ # LATIN SMALL LETTER U WITH DOT BELOW
U+1EE7 ; ủ # LATIN SMALL LETTER U WITH HOOK ABOVE
U+1EE9 ; ứ # LATIN SMALL LETTER U WITH HORN AND ACUTE

U+1EEB ; ừ # LATIN SMALL LETTER U WITH HORN AND GRAVE
U+1EED ; ử # LATIN SMALL LETTER U WITH HORN AND HOOK ABOVE
U+1EEF ; ữ # LATIN SMALL LETTER U WITH HORN AND TILDE
U+1EF1 ; ự # LATIN SMALL LETTER U WITH HORN AND DOT BELOW
U+0076 ; v # LATIN SMALL LETTER V
U+1E7D ; ṽ # LATIN SMALL LETTER V WITH TILDE
U+1E7F ; ṿ # LATIN SMALL LETTER V WITH DOT BELOW
U+1EFD ; ỽ # LATIN SMALL LETTER MIDDLE-WELSH V
U+2C71 ; ⱱ # LATIN SMALL LETTER V WITH RIGHT HOOK
U+2C74 ; ⱴ # LATIN SMALL LETTER V WITH CURL
U+028B ; ʋ # LATIN SMALL LETTER V WITH HOOK
U+028C ; ʌ # LATIN SMALL LETTER TURNED V
U+0077 ; w # LATIN SMALL LETTER W
U+0175 ; ŵ # LATIN SMALL LETTER W WITH CIRCUMFLEX
U+1E81 ; ẁ # LATIN SMALL LETTER W WITH GRAVE
U+1E83 ; ẃ # LATIN SMALL LETTER W WITH ACUTE
U+1E85 ; ẅ # LATIN SMALL LETTER W WITH DIAERESIS
U+1E87 ; ẇ # LATIN SMALL LETTER W WITH DOT ABOVE
U+1E89 ; ẉ # LATIN SMALL LETTER W WITH DOT BELOW
U+1E98 ; ẘ # LATIN SMALL LETTER W WITH RING ABOVE
U+2C73 ; ⱳ # LATIN SMALL LETTER W WITH HOOK
U+01BF ; ƿ # LATIN LETTER WYNN
U+0078 ; x # LATIN SMALL LETTER X
U+1E8B ; ẋ # LATIN SMALL LETTER X WITH DOT ABOVE
U+1E8D ; ẍ # LATIN SMALL LETTER X WITH DIAERESIS
U+0079 ; y # LATIN SMALL LETTER Y
U+00FD ; ý # LATIN SMALL LETTER Y WITH ACUTE
U+00FF ; ÿ # LATIN SMALL LETTER Y WITH DIAERESIS
U+0177 ; ŷ # LATIN SMALL LETTER Y WITH CIRCUMFLEX
U+01B4 ; ƴ # LATIN SMALL LETTER Y WITH HOOK
U+0233 ; ȳ # LATIN SMALL LETTER Y WITH MACRON
U+024F ; ɏ # LATIN SMALL LETTER Y WITH STROKE
U+1E99 ; ẙ # LATIN SMALL LETTER Y WITH RING ABOVE
U+1EF3 ; ỳ # LATIN SMALL LETTER Y WITH GRAVE

U+1EF5 ; ỵ # LATIN SMALL LETTER Y WITH DOT BELOW
U+1EF7 ; ỷ # LATIN SMALL LETTER Y WITH HOOK ABOVE
U+1EF9 ; ỹ # LATIN SMALL LETTER Y WITH TILDE
U+1EFF ; ỿ # LATIN SMALL LETTER Y WITH LOOP
U+1E8F ; ẏ # LATIN SMALL LETTER Y WITH DOT ABOVE
U+021D ; ȝ # LATIN SMALL LETTER YOGH
U+007A ; z # LATIN SMALL LETTER Z
U+017A ; ź # LATIN SMALL LETTER Z WITH ACUTE
U+017C ; ż # LATIN SMALL LETTER Z WITH DOT ABOVE
U+017E ; ž # LATIN SMALL LETTER Z WITH CARON
U+01B6 ; ƶ # LATIN SMALL LETTER Z WITH STROKE
U+0225 ; ȥ # LATIN SMALL LETTER Z WITH HOOK
U+0240 ; ɀ # LATIN SMALL LETTER Z WITH SWASH TAIL
U+1E91 ; ẑ # LATIN SMALL LETTER Z WITH CIRCUMFLEX
U+1E93 ; ẓ # LATIN SMALL LETTER Z WITH DOT BELOW
U+1E95 ; ẕ # LATIN SMALL LETTER Z WITH LINE BELOW
U+2C6C ; ⱬ # LATIN SMALL LETTER Z WITH DESCENDER
U+0292 ; ʒ # LATIN SMALL LETTER EZH
U+01B9 ; ƹ # LATIN SMALL LETTER EZH REVERSED
U+01BA ; ƺ # LATIN SMALL LETTER EZH WITH TAIL
U+01EF ; ǯ # LATIN SMALL LETTER EZH WITH CARON
U+01C0 ; ǀ # LATIN LETTER DENTAL CLICK
U+01C1 ; ǁ # LATIN LETTER LATERAL CLICK
U+01C2 ; ǂ # LATIN LETTER ALVEOLAR CLICK
U+01C3 ; ! # LATIN LETTER RETROFLEX CLICK
U+0294 ; ʔ # LATIN LETTER GLOTTAL STOP

B3. Reserve code points

The following code points are also available for inclusion in Latin script IDNs but the study group only expects them to appear under special circumstances — which do not include use in the root zone of the DNS — and has therefore placed them in a separate table. These code points are listed in numerical order under their Unicode block headings or subheadings:

IPA EXTENSIONS

U+0250 ; ɐ # LATIN SMALL LETTER TURNED A
U+0252 ; ɒ # LATIN SMALL LETTER TURNED ALPHA
U+0258 ; ɘ # LATIN SMALL LETTER REVERSED E
U+025A ; ɚ # LATIN SMALL LETTER SCHWA WITH HOOK
U+025C ; ɜ # LATIN SMALL LETTER REVERSED OPEN E
U+025D ; ɝ # LATIN SMALL LETTER REVERSED OPEN E WITH HOOK
U+025E ; ɞ # LATIN SMALL LETTER CLOSED REVERSED OPEN E
U+0264 ; ɤ # LATIN SMALL LETTER RAMS HORN
U+0265 ; ɥ # LATIN SMALL LETTER TURNED H
U+0267 ; ɧ # LATIN SMALL LETTER HENG WITH HOOK
U+0268 ; ɨ # LATIN SMALL LETTER I WITH STROKE
U+026A ; ɪ # LATIN LETTER SMALL CAPITAL I
U+026B ; ɫ # LATIN SMALL LETTER L WITH MIDDLE TILDE
U+026C ; ɬ # LATIN SMALL LETTER L WITH BELT
U+026D ; ɭ # LATIN SMALL LETTER L WITH RETROFLEX HOOK
U+026E ; ɮ # LATIN SMALL LETTER LEZH
U+026F ; ɯ # LATIN SMALL LETTER TURNED M
U+0270 ; ɰ # LATIN SMALL LETTER TURNED M WITH LONG LEG
U+0271 ; ɱ # LATIN SMALL LETTER M WITH HOOK
U+0273 ; ɳ # LATIN SMALL LETTER N WITH RETROFLEX HOOK
U+0274 ; ɴ # LATIN LETTER SMALL CAPITAL N
U+0276 ; ɶ # LATIN LETTER SMALL CAPITAL OE
U+0277 ; ɷ # LATIN SMALL LETTER CLOSED OMEGA
U+0278 ; ɸ # LATIN SMALL LETTER PHI
U+0279 ; ɹ # LATIN SMALL LETTER TURNED R
U+027A ; ɺ # LATIN SMALL LETTER TURNED R WITH LONG LEG
U+027B ; ɻ # LATIN SMALL LETTER TURNED R WITH HOOK
U+027C ; ɼ # LATIN SMALL LETTER R WITH LONG LEG
U+027E ; ɾ # LATIN SMALL LETTER R WITH FISHHOOK
U+0280 ; ʀ # LATIN LETTER SMALL CAPITAL R
U+0281 ; ʁ # LATIN LETTER SMALL CAPITAL INVERTED R
U+0282 ; ʂ # LATIN SMALL LETTER S WITH HOOK
U+0284 ; ʄ # LATIN SMALL LETTER DOTLESS J WITH STROKE AND HOOK

U+0285 ; ʅ # LATIN SMALL LETTER SQUAT REVERSED ESH
U+0286 ; ʆ # LATIN SMALL LETTER ESH WITH CURL
U+0287 ; ʇ # LATIN SMALL LETTER TURNED T
U+0289 ; ʉ # LATIN SMALL LETTER U BAR
U+028A ; ʊ # LATIN SMALL LETTER UPSILON
U+028D ; ʍ # LATIN SMALL LETTER TURNED W
U+028E ; ʎ # LATIN SMALL LETTER TURNED Y
U+028F ; ʏ # LATIN LETTER SMALL CAPITAL Y
U+0290 ; ʐ # LATIN SMALL LETTER Z WITH RETROFLEX HOOK
U+0291 ; ʑ # LATIN SMALL LETTER Z WITH CURL
U+0293 ; ʓ # LATIN SMALL LETTER EZH WITH CURL
U+0295 ; ʕ # LATIN LETTER PHARYNGEAL VOICED FRICATIVE
U+0296 ; ʖ # LATIN LETTER INVERTED GLOTTAL STOP
U+0297 ; ʗ # LATIN LETTER STRETCHED C
U+0298 ; ʘ # LATIN LETTER BILABIAL CLICK
U+0299 ; ʙ # LATIN LETTER SMALL CAPITAL B
U+029A ; ʚ # LATIN SMALL LETTER CLOSED OPEN E
U+029B ; ʛ # LATIN LETTER SMALL CAPITAL G WITH HOOK
U+029C ; ʜ # LATIN LETTER SMALL CAPITAL H
U+029D ; ʝ # LATIN SMALL LETTER J WITH CROSSED-TAIL
U+029E ; ʞ # LATIN SMALL LETTER TURNED K
U+029F ; ʟ # LATIN LETTER SMALL CAPITAL L
U+02A0 ; ʠ # LATIN SMALL LETTER Q WITH HOOK
U+02A1 ; ʡ # LATIN LETTER GLOTTAL STOP WITH STROKE
U+02A2 ; ʢ # LATIN LETTER REVERSED GLOTTAL STOP WITH STROKE
U+02A3 ; ʣ # LATIN SMALL LETTER DZ DIGRAPH
U+02A4 ; ʤ # LATIN SMALL LETTER DEZH DIGRAPH
U+02A5 ; ʥ # LATIN SMALL LETTER DZ DIGRAPH WITH CURL
U+02A6 ; ʦ # LATIN SMALL LETTER TS DIGRAPH
U+02A7 ; ʧ # LATIN SMALL LETTER TESH DIGRAPH
U+02A8 ; ʨ # LATIN SMALL LETTER TC DIGRAPH WITH CURL
U+02A9 ; ʩ # LATIN SMALL LETTER FENG DIGRAPH
U+02AA ; ʪ # LATIN SMALL LETTER LS DIGRAPH
U+02AB ; ʫ # LATIN SMALL LETTER LZ DIGRAPH

U+02AC ; ʬ # LATIN LETTER BILABIAL PERCUSSIVE

U+02AD ; ʭ # LATIN LETTER BIDENTAL PERCUSSIVE

PHONETIC EXTENSIONS

U+1D6B ; ᵫ # LATIN SMALL LETTER UE

U+1D6C ; ᵬ # LATIN SMALL LETTER B WITH MIDDLE TILDE

U+1D6D ; ᵭ # LATIN SMALL LETTER D WITH MIDDLE TILDE

U+1D6E ; ᵮ # LATIN SMALL LETTER F WITH MIDDLE TILDE

U+1D6F ; ᵯ # LATIN SMALL LETTER M WITH MIDDLE TILDE

U+1D70 ; ᵰ # LATIN SMALL LETTER N WITH MIDDLE TILDE

U+1D71 ; ᵱ # LATIN SMALL LETTER P WITH MIDDLE TILDE

U+1D72 ; ᵲ # LATIN SMALL LETTER R WITH MIDDLE TILDE

U+1D73 ; ᵳ # LATIN SMALL LETTER R WITH FISHHOOK AND MIDDLE TILDE

U+1D74 ; ᵴ # LATIN SMALL LETTER S WITH MIDDLE TILDE

U+1D75 ; ᵵ # LATIN SMALL LETTER T WITH MIDDLE TILDE

U+1D76 ; ᵶ # LATIN SMALL LETTER Z WITH MIDDLE TILDE

U+1D79 ; ᵹ # LATIN SMALL LETTER INSULAR G

U+1D7A ; ᵺ # LATIN SMALL LETTER TH WITH STRIKETHROUGH

U+1D7B ; ᵻ # LATIN SMALL LETTER CAPITAL LETTER I WITH STROKE

U+1D7C ; ᵼ # LATIN SMALL LETTER IOTA WITH STROKE

U+1D7D ; ᵽ # LATIN SMALL LETTER P WITH STROKE

U+1D7E ; ᵾ # LATIN SMALL LETTER CAPITAL LETTER U WITH STROKE

U+1D7F ; ᵿ # LATIN SMALL LETTER UPSILON WITH STROKE

U+1D80 ; ᶀ # LATIN SMALL LETTER B WITH PALATAL HOOK

U+1D81 ; ᶁ # LATIN SMALL LETTER D WITH PALATAL HOOK

U+1D82 ; ᶂ # LATIN SMALL LETTER F WITH PALATAL HOOK

U+1D83 ; ᶃ # LATIN SMALL LETTER G WITH PALATAL HOOK

U+1D84 ; ᶄ # LATIN SMALL LETTER K WITH PALATAL HOOK

U+1D85 ; ᶅ # LATIN SMALL LETTER L WITH PALATAL HOOK

U+1D86 ; ᶆ # LATIN SMALL LETTER M WITH PALATAL HOOK

U+1D87 ; ᶇ # LATIN SMALL LETTER N WITH PALATAL HOOK

U+1D88 ; ᶈ # LATIN SMALL LETTER P WITH PALATAL HOOK

U+1D89 ; ᶉ # LATIN SMALL LETTER R WITH PALATAL HOOK

U+1D8A ; ᶊ # LATIN SMALL LETTER S WITH PALATAL HOOK

U+1D8B ; ʃ᷉ # LATIN SMALL LETTER ESH WITH PALATAL HOOK
U+1D8C ; ᶌ # LATIN SMALL LETTER V WITH PALATAL HOOK
U+1D8D ; x᷉ # LATIN SMALL LETTER X WITH PALATAL HOOK
U+1D8E ; ʒ᷉ # LATIN SMALL LETTER Z WITH PALATAL HOOK
U+1D8F ; aᶑ # LATIN SMALL LETTER A WITH RETROFLEX HOOK
U+1D90 ; ɑᶑ # LATIN SMALL LETTER ALPHA WITH RETROFLEX HOOK
U+1D91 ; dᶑ # LATIN SMALL LETTER D WITH HOOK AND TAIL
U+1D92 ; eᶑ # LATIN SMALL LETTER E WITH RETROFLEX HOOK
U+1D93 ; ɛᶑ # LATIN SMALL LETTER OPEN E WITH RETROFLEX HOOK
U+1D94 ; ɜᶑ # LATIN SMALL LETTER REVERSED OPEN E WITH RETROFLEX HOOK
U+1D95 ; əᶑ # LATIN SMALL LETTER SCHWA WITH RETROFLEX HOOK
U+1D96 ; iᶑ # LATIN SMALL LETTER I WITH RETROFLEX HOOK
U+1D97 ; ɔᶑ # LATIN SMALL LETTER OPEN O WITH RETROFLEX HOOK
U+1D98 ; ʃᶑ # LATIN SMALL LETTER ESH WITH RETROFLEX HOOK
U+1D99 ; uᶑ # LATIN SMALL LETTER U WITH RETROFLEX HOOK
U+1D9A ; ʒᶑ # LATIN SMALL LETTER EZH WITH RETROFLEX HOOK

LATIN EXTENDED-C

Claudian letters

U+2C76 ; ⱶ # LATIN SMALL LETTER HALF H

Additions for UPA

U+2C77 ; ⱷ # LATIN SMALL LETTER TAILLESS PHI
U+2C78 ; ⱸ # LATIN SMALL LETTER E WITH NOTCH
U+2C79 ; ⱹ # LATIN SMALL LETTER TURNED R WITH TAIL
U+2C7A ; ⱺ # LATIN SMALL LETTER O WITH LOW RING INSIDE
U+2C7B ; ⱻ # LATIN LETTER SMALL CAPITAL TURNED E

LATIN EXTENDED-D

Egyptological Additions

U+A723 ; ꜣ # LATIN SMALL LETTER EGYPTOLOGICAL ALEF
U+A725 ; ꜥ # LATIN SMALL LETTER EGYPTOLOGICAL AIN

Mayanist Additions

U+A727 ; ꜧ # LATIN SMALL LETTER HENG
U+A729 ; ꜩ # LATIN SMALL LETTER TZ
U+A72B ; ꜫ # LATIN SMALL LETTER TRESILLO
U+A72D ; ꜭ # LATIN SMALL LETTER CUATRILLO
U+A72F ; ꜯ # LATIN SMALL LETTER CUATRILLO WITH COMMA

Medievalist Additions

U+A730 ; ꜰ # LATIN LETTER SMALL CAPITAL F
U+A731 ; ꜱ # LATIN LETTER SMALL CAPITAL S
U+A733 ; ꜳ # LATIN SMALL LETTER AA
U+A735 ; ꜵ # LATIN SMALL LETTER AO
U+A737 ; ꜷ # LATIN SMALL LETTER AU
U+A739 ; ꜹ # LATIN SMALL LETTER AV
U+A73B ; ꜻ # LATIN SMALL LETTER AV WITH HORIZONTAL BAR
U+A73D ; ꜽ # LATIN SMALL LETTER AY
U+A73F ; ꜿ # LATIN SMALL LETTER REVERSED C WITH DOT
U+A741 ; ꝁ # LATIN SMALL LETTER K WITH STROKE
U+A743 ; ꝃ # LATIN SMALL LETTER K WITH DIAGONAL STROKE
U+A745 ; ꝅ # LATIN SMALL LETTER K WITH STROKE AND DIAGONAL
U+A747 ; ꝇ # LATIN SMALL LETTER BROKEN L
U+A749 ; ꝉ # LATIN SMALL LETTER L WITH HIGH STROKE
U+A74B ; ꝋ # LATIN SMALL LETTER O WITH LONG STROKE OVERLAY
U+A74D ; ꝍ # LATIN SMALL LETTER O WITH LOOP
U+A74F ; ꝏ # LATIN SMALL LETTER OO
U+A751 ; ꝑ # LATIN SMALL LETTER P WITH STROKE THROUGH DES
U+A753 ; ꝓ # LATIN SMALL LETTER P WITH FLOURISH
U+A755 ; ꝕ # LATIN SMALL LETTER P WITH SQUIRREL TAIL
U+A757 ; ꝗ # LATIN SMALL LETTER Q WITH STROKE THROUGH DES
U+A759 ; ꝙ # LATIN SMALL LETTER Q WITH DIAGONAL STROKE
U+A75B ; ꝛ # LATIN SMALL LETTER R ROTUNDA
U+A75D ; ꝝ # LATIN SMALL LETTER RUM ROTUNDA
U+A75F ; ꝟ # LATIN SMALL LETTER V WITH DIAGONAL STROKE

U+A761 ; ꝡ # LATIN SMALL LETTER VY

U+A763 ; ꝣ # LATIN SMALL LETTER VISIGOTHIC Z

U+A765 ; ꝥ # LATIN SMALL LETTER THORN WITH STROKE

U+A767 ; ꝧ # LATIN SMALL LETTER THORN WITH STROKE THROUGH

U+A769 ; ꝩ # LATIN SMALL LETTER VEND

U+A76B ; ꝫ # LATIN SMALL LETTER ET

U+A76D ; ꝭ # LATIN SMALL LETTER IS

U+A76F ; ꝯ # LATIN SMALL LETTER CON

U+A771 ; ꝱ # LATIN SMALL LETTER DUM

U+A772 ; ꝲ # LATIN SMALL LETTER LUM

U+A773 ; ꝳ # LATIN SMALL LETTER MUM

U+A774 ; ꝴ # LATIN SMALL LETTER NUM

U+A775 ; ꝵ # LATIN SMALL LETTER RUM

U+A776 ; ꝶ # LATIN LETTER SMALL CAPITAL RUM

U+A777 ; ꝷ # LATIN SMALL LETTER TUM

U+A778 ; ꝸ # LATIN SMALL LETTER UM

Insular and Celticist letters

U+A77A ; ꝺ # LATIN SMALL LETTER INSULAR D

U+A77C ; ꝼ # LATIN SMALL LETTER INSULAR F

U+A77F ; ꝿ # LATIN SMALL LETTER TURNED INSULAR G

U+A781 ; ꞁ # LATIN SMALL LETTER TURNED L

U+A783 ; ꞃ # LATIN SMALL LETTER INSULAR R

U+A785 ; ꞅ # LATIN SMALL LETTER INSULAR S

U+A787 ; ꞇ # LATIN SMALL LETTER INSULAR T

Orthographic letters for glottals

U+A78C ; ' # LATIN SMALL LETTER SALTILLO

Ancient Roman epigraphic letters

U+A7FB ; Ꟈ # LATIN EPIGRAPHIC LETTER REVERSED F

U+A7FC ; Ꟊ # LATIN EPIGRAPHIC LETTER REVERSED P

U+A7FD ; Ꟍ # LATIN EPIGRAPHIC LETTER INVERTED M

U+A7FE ; Ꟈ # LATIN EPIGRAPHIC LETTER I LONGA
U+A7FF ; Ꟙ # LATIN EPIGRAPHIC LETTER ARCHAIC M

Phonetic symbol

U+A78E ; ꞎ # LATIN SMALL LETTER L WITH RETROFLEX HOOK AND BELT

Janalif letters

U+A791 ; ꞑ # LATIN SMALL LETTER N WITH DESCENDER

Latvian letters for pre-1921 orthography

U+A7A1 ; ꞡ # LATIN SMALL LETTER G WITH OBLIQUE STROKE
U+A7A3 ; ꞣ # LATIN SMALL LETTER K WITH OBLIQUE STROKE
U+A7A5 ; ꞥ # LATIN SMALL LETTER N WITH OBLIQUE STROKE
U+A7A7 ; ꞧ # LATIN SMALL LETTER R WITH OBLIQUE STROKE
U+A7A9 ; ꞩ # LATIN SMALL LETTER S WITH OBLIQUE STROKE

Addition for UPA

U+A7FA ; ꟻ # LATIN LETTER SMALL CAPITAL TURNED M