# Principles for Inclusion and Exclusion of Code Points for Latin Script (Latin Generation Panel) for the Root Zone

This is a working document which captures the principles for inclusion and exclusion decisions regarding code points for Latin Script as part of the Label Generation Rule-set (LGR) for the Root zone. The principles are guidelines, which may evolve as additional analysis is undertaken during the course of the work.

The principles for inclusion and exclusion of codes points are listed in their respective category.

# Definitions

**Language**, using Latin Script, is defined as any natural human language, which is written in Latin Script, as a writing system. The relative status of language (i.e., political or legal status and absolute or relative frequency of use of language) is not considered for the current purposes. Accordingly, notions such as dialect and regiolect are all considered equivalent to language, while notions such as official language, national language, standard language and vernacular, are not considered at all.

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

**Mark Code Point** is a Unicode code point with General Category property value of M*x* (Mn, Mc, Me), as defined in the Unicode Character Database. (See Appendix A)

**Established contemporary use** of a letter is its active use in educational resources, published material, media, etc. by diverse sources, irrespective of hand written or typed, within a community. Other factors to consider are the age group using the code point (children, young adults, adults, elderly) and the diversity of users of the code point (individual, individuals, sub-community or community at large). There may be multiple sources for acquiring such evidence including (but not limited to) the following: (i) Language community, (ii) Members of Latin GP, (iii) Other expert, (iv) Language Table submitted by ccTLD in the context of IDNA2008 in the IANA repository, and (v) Published standard (e.g. by a language authority, other national body, international body, etc.).

# Inclusion Principles

At least one of the principles must apply. If a code point is included and delegated as part of the label, the code point cannot be retracted. Therefore, applicable rule in this category must be definitive, while remaining conservative to include only the code points necessary for the root zone.

# *2.1 Language has level of 0-4 on EGID Scale*

# *2.2 Letter code point* which is a *letter* and has *established contemporary use* in a language

# *2.3 Code Point which is vowel consonants with context (Code point + Mark), Language reference(after some character) , Sequence, modified letters Mark code point* which represents a *required mark*, where at least one of the *letters* it forms has *established contemporary use* in a language. Mark and modified letter shall be represented in sequence. Lower letter or modified letter before or after. Marks does not have any space.

# Exclusion Principles

At least one of the principles must apply. If a code point is excluded, it is not easily possible to include it in the next versions of the LGR. Therefore, applicability of the principle in this category must be definitive. If the principle is not definitively applicable for a code point, the code point should be considered for deferral.

3.1 Code point DISALLOWED by IDNA 2008 protocol

3.2 Code point presents a security or stability issue which cannot be resolved at any other stage of the analysis (i.e., stage of determining code points, variants or whole label rules)

3.3 Code point not listed in the **Latin** GP proposal

3.4Code point either deprecated or not recommended for use in Unicode Standard; exception being it meets one of the inclusion criteria with no alternative code point(s)

3.5 Code point for historic use with no *established contemporary use*

*3.6* Code point representing a technical sign, such as encountered in religious texts, which is not otherwise used as described in Section 2 above.

3.7 Code point does not meet any of the inclusion criteria and is **only used** for other purposes, for example:

# An optional mark

# A formatting character or mark

# A numerical digit

# A punctuation mark

# An honorific mark or symbol

# A mathematical symbol

# Appendix A

# Details of Character Properties

Lu = Letter, uppercase

Ll = Letter, lowercase

Lt = Letter, titlecase

Lm = Letter, modifier

Lo = Letter, other

Mn = Mark, nonspacing

Mc = Mark, spacing combining

Me = Mark, enclosing

Nd = Number, decimal digit

Nl = Number, letter

No = Number, other

Pc = Punctuation, connector

Pd = Punctuation, dash

Ps = Punctuation, open

Pe = Punctuation, close

Pi = Punctuation, initial quote (may behave like Ps or Pe depending on usage)

Pf = Punctuation, final quote (may behave like Ps or Pe depending on usage)

Po = Punctuation, other

Sm = Symbol, math

Sc = Symbol, currency

Sk = Symbol, modifier

So = Symbol, other

Zs = Separator, space

Zl = Separator, line

Zp = Separator, paragraph

Cc = Other, control

Cf = Other, format

Cs = Other, surrogate

Co = Other, private use

Cn = Other, not assigned (including non-characters)

Source for Unicode tables / charts : <http://www.unicode.org/charts/>