Principles for Inclusion and Exclusion of Code Points in Latin Script for the Root Zone (Latin LGR)

Draft version - Please send feedback to LatinGP@icann.org *by 23 October 2017*

This document defines the principles for inclusion and exclusion decisions regarding Code Points for Latin Script as part of the Label Generation Rules (LGR) for the Root Zone. The principles 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 below.

1. Definitions

Language: The present document and its principles deal with any language making use of Latin script[1] today. Languages are restricted to natural human languages in active use. Both the socio-political situation (such as the political or legal status of a language in a country or community) and the socio-linguistic roles of languages in society (such as the absolute or relative frequency of use) are explicitly not considered for the current purposes. Super- or sub-units of languages, such as dialect, regiolect (a dialect spoken in a particular geographical region), or language clusters, are all considered equivalent to language. However notions such as official language, national language, standard language and vernacular, are not considered at all in determining whether something is a language.

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

Code Point Sequence is a sequence of two or more Code Points (e.g. Letter Code Point followed by a Mark Code Point).

Established contemporary use of a letter means it is in active use by a community today. Such use may be demonstrated by, for example, educational resources, published material, media, or other materials and sources. This does not depend on their material or non-material form, such as handwritten or typed manuscripts or digitally produced text. There

[1] Latin script is also known as Roman script in academic literature.

may be multiple sources for acquiring such evidence, including (but not limited to) the following:

- Members of Language communities,
- Members of the Latin GP,
- Other experts
- Language tables submitted by ccTLD in the context of IDNA 2008 in the IANA repository, and
- Published standards (e.g. by a language authority or any other national or international body).

2. Inclusion Principles

If a Code Point is included and delegated as part of the label, the Code Point cannot be retracted in future revisions of the LGR. All applicable criteria must be met to include a Code Point.

1. Only languages which have a rating of levels of 0-4 under the Expanded Graded Intergenerational Disruption Scale (EGIDS) are considered as supporting the inclusion of a Code Point. Languages with EGIDS 5 may be included in special cases where there is additional evidence that it is in widespread use, notwithstanding its formal EGIDS rating.
2. Code Points may only be included if they have established contemporary use in one or more of the languages considered.
3. If the Code Point in question is a Mark Code Point, then it can only be included in its context. That is, a Mark Code Point is included as part of a sequence consisting of a Lower Letter (Ll) or Other Letter (Lo) and the subsequent mark or marks.
4. Any combination of Code Points is defined by its sequence. To be included, a sequence must be supported by some included language in the same way as a separate Code Point of type Ll or Lo.
5. If a character can be represented by multiple Code Point Sequences, each Code Point Sequence must be separately justified to be included.
6. A Code Point Sequence can only be included if there is no pre-composed alternative available, unless there is specific evidence that a language eligible for inclusion under Criteria 1 makes alternate use of such a sequence.
7. If the Code Point in question is a Modifier letter (Lm), then it can only be included together with its context. That is a sequence of Lm plus Ll or Lo (or the other way around), unless there is strong evidence that the Lm can be used in any context, or that such a sequence or order cannot be defined.

3. Exclusion Principles

A Code Point is excluded if at least one of these exclusion principles is met. If a Code Point can neither be included nor excluded on the basis of these principles, the Code Point is automatically excluded from the proposed LGR for Latin Script, per RFC 6912.

1. The Code Point is DISALLOWED or UNASSIGNED by IDNA 2008 protocol.
2. The Code Point presents a security or stability issue which cannot be resolved at any other stage of the analysis (e.g., stage of determining Code Points, variants, Contextual Rules or Whole Label Evaluation Rules).
3. The Code Point is either deprecated or not recommended for use in Unicode Standard -- unless it meets all of the applicable inclusion criteria, with no alternative Code Point or Code Point sequence.
4. The Code Point is used exclusively in a subset of textual genres, such as technical or religious texts, and is not otherwise used as described in Section 2 above.
5. The Code Point is predominantly used in one of the following functions, apart from any other uses in orthography**:**
    a. Formatting character or mark
    b. Numerical digit
    c. Punctuation mark
    d. Honorific mark or symbol
    e. 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, non-spacing

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/