Principles for Inclusion, Exclusion or Deferral of Code Points for Arabic Script LGR for the Root Zone

Task Force on Arabic Internationalized Domain Names (TF-AIDN)

This is a working document which captures the high level principles for inclusion, exclusion and deferral of decisions regarding code points for Arabic Script as part of the Label Generation Rule-set (LGR) for the Root zone. The principles are high-level guidelines, which may evolve as additional analysis is undertaken during the course of the work. Further, as they present general guidelines and not strict rules and definitions, a principle may overlap or disagree from other principles. Where a principle overlaps, it reinforces the other. Where two principles disagree, the eventual decision for relevant code points will have to be done on a case-to-case basis.

The principles for inclusion, exclusion and deferral of codes points are listed in their respective category. Though some care has been taken to avoid duplication, however, the contents of the sections may still not be mutually exclusive.

1. Definitions

Language, using Arabic script, is defined as any natural human language, which is written in Arabic 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, for Arabic script, is a grapheme representing a combination of one Rasm and zero or more Ijam, representing a consonantal or vocalic sound, in whole or partially.

Required Mark, for an Arabic-script based language, is a grapheme required to form a letter, e.g. most of Ijam marks.

Optional Mark, for an Arabic-script based language, is a grapheme optionally used and is not required to form a letter, e.g. some Tashkil marks.

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 TF-AIDN, (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.).

2. 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. If the rule is not definitively applicable for a code point, the code point should be considered for deferral.

2.1. *Letter code point* which is a *letter* and has *established contemporary use* in a language

2.2. *Mark code point* which represents a *required mark*, where at least one of the *letters* it forms has *established contemporary use* in a language

2.3. Code point which represents a combination of *letters* in a language which has *established contemporary use*, where at least one of the constituent letters cannot be represented by a combination of *letter code points* and *mark code points*.

2.4. Code point which represents a *lexical word* or *phrase* in a language, which has *established contemporary use* and cannot be decomposed into a sequence of code points representing *letter code points* and *mark code points*.

3. 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 Arabic GP proposal

3.4. Code 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 specifically 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:

a. An optional mark
b. A formatting character or mark
c. A numerical digit
d. A punctuation mark
e. An honorific mark or symbol
f. A mathematical symbol

4. Deferral Principles

This category has been developed to balance the conservatism for LGR inclusion and the definitiveness of the exclusion. However, the set of code points deferred should be kept to a minimum, as there is a chance that such code point cannot be included at a later stage because it may have adverse stability impact on already included code point or (worse) on delegated labels.

4.1. Code point which can neither be confirmed for inclusion nor for exclusion based on principles in Sections 2 and 3 above. For example:

a. Code point possibly meets one of the inclusion criteria, but contemporary use cannot be established
b. Code point meets one of the inclusion criteria except that it is not currently used, but has been used in writing by a community recently and its use may still be revitalized

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: Table 4-9 at http://www.unicode.org/versions/Unicode6.2.0/ch04.pdf.