Guidelines for IDN Reference Tables

.SE (The Internet Infrastructure Foundation)
https://www.iis.se/en/

2014-10-10
Version A

1. Introduction

The purpose of this document is to be guidelines for the production of IDN Reference Tables for Pre-Delegation Testing (PDT) for the ICANN New gTLD program. The IDN Reference Tables are produced by .SE as the PDT Service Provider (https://github.com/dotse/IDN-ref-tables).

The main purpose of these tables is to be used as reference tables for IDN testing for PDT. They can also be used as reference tables for any IDN registration. They are not supposed to be used for any other used but IDN.

These tables set the maximum repertoire of language and script tables, a more restricted table is accepted by this framework. The reference tables will be updated if needed.

2. Variants

“Variant” is a term that has been used to indicate some sort of relationship between two or more labels or names that are to be considered to be the same or to be confusable. Many IDN tables used by various TLD supports the concept of variants that put limitations on labels that are considered to be variants. The IDN Reference Tables do not cover the question of variants. Further discussions on concept can be found in https://www.icann.org/en/system/files/files/idn-vip-integrated-issues-final-clean-20feb12-en.pdf.

3. Only valid code points

An IDN table must only contain IDNA 2008 valid code points. The IDNA protocol defines in RFC 5892 four categories that all Unicode code points fall into:

- Protocol Valid (PVALID)
- Contextual Rule Required (CONTEXTO/CONTEXTJ)
- Disallowed
- Unassigned

For an extensive list of what code points that fall into which of the above categories see http://www.iana.org/assignments/idna-tables.

Only code points that are PVALID, CONTEXTO or CONTEXTJ may appear in an IDN table (RFC 5891).

The border between PVALID and UNASSIGNED depends on the Unicode version. Code points may go from UNASSIGNED to PVALID (but never the other way). The IDN Reference tables are based on version 6.3.0 of the Unicode standard, currently being the latest version of the Unicode standard for which the "IDNA Derived Properties" have been officially derived (http://www.iana.org/assignments/idna-tables).

Code points that are DISALLOWED based on one version of Unicode will always remain disallowed. An unassigned code point may go to DISALLOWED in a later or future version of Unicode.

4. Metadata

An IDN table should contain metadata stating what language (if language table) or script (if script table) it is designed for, using the format given in RFC 5646. It should also state what Unicode version it is based on. Version of the table, date of creation or update and responsible for the table should also be included.

5. General rule -- only one explicit Unicode script

There are two categories of Unicode script property values:

- Explicit script property value (e.g. Latin, Cyrillic or Han)
- Special script property value (i.e. Common or Inherited)

As a general rule, all code points in an IDN table must have the same script property value. Other code points in the table may have a special script property value (i.e. Common or Inherited). More about Common and Inherited code points and exceptions to the general rule below.

6. Special rule for CJK IDN tables

IDN tables classified as Chinese language, Japanese language or Korean language may have code points with different explicit script property values. The following is valid for IDN tables for the three languages:

- Chinese language: Latin and Han scripts
- Japanese language: Latin, Hiragana, Katakana and Han scripts
- Korean language: Latin, Hangul and Han scripts

The writing above does not mean that an IDN table can contain any code point from the listed scripts. For each language, only the relevant code points may be included (more about language specific rules below).

An IDN table labeled as a script table can never have code points with more than one Unicode script property (i.e. the script property in question) except for code points having Common and Inherited property, as explained below.

7. General rule -- DH code points in any IDN table

As a general rule, all IDN tables may contain the DH (digit and hyphen) code points:

- European digits 0030..0039
- Hyphen U+002D

Depending on other code points in the table, contextual rules the European digits may be necessary. Having contextual rules for hyphen is always necessary. See more below on contextual rules.

8. Common and Inherited code points can be included if relevant

Besides the DH cluster, an IDN table can contain other common code points or inherited code points if relevant for the language (if language table) or script (if script table). E.g. in an Arabic script table Arabic-Indic digits (0660..0669) are relevant, but not in a Latin script table.

The Unicode database contains the file ScriptExtensions.txt that should be consulted to determine with what explicit Unicode scripts that a specific COMMON code point can be used with. Also UAX#24 should be consulted.

9. Contextual rules

Many code points require contextual rules, and those rules must be included in the IDN table. There are different groups of code points that require contextual rules:

- Rules for HYPHEN-MINUS (U+002D) as specified in RFC5891.
- Rules for code points belonging to general category "mark" (nonspacing mark, spacing mark or enclosing mark) in the Unicode database as specified in RFC5891 and UAX#24 (e.g. U+0308, U+1B40).

- Rules for code points belonging to general category "modifier letters" the Unicode database (e.g. U+16F99).
- Rules for CONTEXTO and CONTEXTJ code points as specified in RFC5891 (e.g. U+ 0660).
- Specific rules for RTL labels specified in RFC5893 (e.g. U+0030).
- Rules for COMMON and INHERITED code points found in UAX#24 and in the Unicode database (e.g. U+30FC).
- Other code points that the Unicode database specifies should have contextural rules (e.g. U+0E40).

Rules should be explicitly stated in the IDN table.

10. Specific considerations for language tables

A language table shall contain only code points that are used for the writing of that language. If the writing tradition needs code points from different explicit Unicode scripts it is permissible to include those in an IDN language table.

- The code points in the table should match the writing tradition of that language of today. The writing tradition used by official institutions and major publishers in that language should be especially considered. Code points only used in past times should not be included.
- If the language has an established alphabet or equivalent where the characters of the language are listed.
- Code points that are used in words of the language can be included even if the code points are very rare, but in contemporary use.
- Code points only used in (proper) names should not be included since names usually transcend the context of one specific language.
- If the language has different writing traditions based on different explicit Unicode scripts, e.g. Latin versus Cyrillic, and the writing is done based on one but not both scripts, then multiple IDN language tables can be created, but characters from different scripts should not be mixed in the same table.
- If the language has a writing tradition based on different explicit Unicode scripts and the code points are mixed, then all these code points from the different scripts can be mixed in the same table. E.g. Japanese with Hiragana, Katakana, Han and Latin.
- If the IDN table contains some code point of category "mark" (non-spacing mark, spacing mark or enclosing mark), e.g. U+030A, or code point of category "modifier letter", e.g. U+16F99, there must be rules regulating when that could be used.

Determination of what is the writing tradition of a specific language should be based on established and published sources.

11. Specific considerations for script table

A script table must only contain code points for one explicit Unicode script. Relevant COMMON or INHERITED code points may also be included. If the IDN table contains some code point of category "mark" (non-spacing mark, spacing mark or enclosing mark), e.g. U+030A, or code point of category "modifier letter", e.g. U+16F99, there must be rules regulating when that could be used.

12. References

- IDN Reference Tables, https://github.com/dotse/IDN-ref-tables
- RFC 5646, " Tags for Identifying Languages", https://tools.ietf.org/html/rfc5646
- RFC 5891, "Internationalized Domain Names in Applications (IDNA): Protocol", http://tools.ietf.org/html/rfc5891
- RFC 5892, " The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)", http://tools.ietf.org/html/rfc5892
- RFC 5893: “Right-to-Left Scripts for Internationalized Domain Names for Applications (IDNA)” (proposed standard): http://tools.ietf.org/html/rfc5893
- Unicode database version 6.3.0, http://www.unicode.org/Public/6.3.0/ucd/
- UAX#24 (Unicode Standard Annex #24), "Unicode Script Property": http://www.unicode.org/reports/tr24
- "The IDN Variant Issues Project: A Study of Issues Related to the Management of IDN Variant TLDs", https://www.icann.org/en/topics/idn/idn-vip-integrated-issues-final-clean-20feb12-en.pdf
- "IDNA Parameters", http://www.iana.org/assignments/idna-tables