Reference Label Generation Rules (LGR) for the Second Level — Overview and Summary

REVISION – May 20, 2016

Table of Contents

1 Overview

This document describes a set of proposed Reference Label Generation Rules (LGR) for the Second Level. These language-based LGRs were developed according to the "Guidelines for Developing Reference LGRs for the Second Level" [Guidelines]. The guidelines define a process that builds on the results of a previous project [IIS] but provides additional review and development, documentation and translation to XML [XML-LGR]. In some cases these LGR extend the repertoire compared to [IIS].

The LGRs are specific to a given language (and in some cases the combination of language and script) but not necessarily specific to a given user community. Each file has been reviewed by one or more linguistic experts, as well reviewed by a separate expert for DNS stability and security issues. The result of this development work is presented here for public comment.

The reader of this document is assumed to be familiar with the [Guidelines].

1.1 Reference Label Generation Rules (LGR) Files

The normative definition of each reference LGR is provided as an XML file, available at https://www.icann.org/resources/pages/second-level-lgr-2015-06-21-en.

The Label Generation rules are expressed using a standard format defined in "Representing Label Generation Rulesets in XML" [XML-LGR].

Each of these files contains all the Label Generation Rules applicable to labels from that language, and only those rules. Each file contains a complete description, a repertoire with optional variants, and WLE Rules, as well as detailed references that link each included code point to a reference providing data for justifying its inclusion.

From each XML file, a non-normative HTML presentation is generated mechanically, also available at https://www.icann.org/resources/pages/second-level-lgr-2015-06-21-en. These are provided for ease of review. The HTML presentation is augmented by summary data as well as data extracted from the Unicode Character Database [UCD].

2 Notes

The development and review process followed the [Guidelines]. The following notes provide some additional highlights as well as information that is not specific to an individual file.

2.1 Repertoire

The repertoire for each LGR is based on a consensus repertoire derived from the sources consulted. In many cases this caters to more than the code points needed to write the native vocabulary of the language, by including code points that are in common use for loan words and the like. Where a language has multiple user communities with some variation of usage, a single, combined LGR was produced. The details are described in each of the LGRs.

2.1.1 Sources for Repertoire

In determining the repertoire a large number of sources was investigated, from spelling dictionaries released by language authorities, RCFs and national or international standards, to other sources such as ordinary dictionaries, the Common Locale Data Repository (a project of the Unicode Consortium) [CLDR] and finally existing IDN practice for ccTLDs aimed at users of a native language. The sources and their contribution to the development of the repertoire are documented in detail in each of the LGRs.

2.2 Extended Code Points

Many, though not all, of the languages are written by compact communities, that are in contact with other languages in the same region or in the same country. In those cases, native users may have familiarity with or need for access to an extended set of code points, for example for names of people or places. The Reference LGRs provide for those code points, if they aren’t already catered for in the core repertoire, by listing them as “extended-cp”. As written, the LGRs treat these extended code points as ineligible for a label, but users could easily remove the restriction to tailor the LGR to their needs.

This is in contrast to script-based LGRs that typically provide for the full repertoire needed for all languages sharing a common script, or those country-based LGRs, that provide for the needs of users from the same country or territory, irrespective of whether they write a majority or minority language prevalent in the country.

The language-based reference LGRs provided here could be used as "building blocks" in assembling local, regional or script-based LGRs. When used in that fashion, care must be taken that the resulting LGR provides for a consistent treatment of variants, for example. This process is called *integration,* and the details depend on the actual set of LGRs to be combined. The current project does not attempt to "pre-integrate" these language-based LGRs.

2.3 Excluded Code Points

For most languages, some sources include a large number of very rarely used code points or some that are historic or limited to special purposes, like poetry and religious works. Such uses are rarely germane to IDNs, but for ease of public review, all of these code points are explicitly listed in the LGRs as "excluded". Doing so facilitates both the reviewers' task of determining that nothing critical was left out, but also makes it easier to adjust the LGR in response to comments. After the round of public comments, any remaining excluded code points will be removed completely.

2.4 Sequences

In a small number of cases, code points occur only in fixed combinations. Where that is the case, the repertoire contains these code points only as part of explicitly specified code point sequence. This prevents unneeded combinations.

2.5 Variants

The majority of language based LGRs does not include the definition of any variants. Where variants are included, their selection was informed by existing registry practice, but also by the work performed at ICANN on the script LGRs for the Root Zone.

2.6 Whole Label Evaluation (WLE) Rules

WLE rules implement a further constraint on labels, for example, by limiting which code points can occur at the beginning of a label, or whether certain code points may show up simultaneously in the same label. Context rules are a form of WLE rule that defines a constraint on the surrounding context for a given code point (see [XML-LGR]).

Because the XML format for the LGR supports machine-evaluation of labels for validity, these reference LGRs include all relevant constraints on labels defined in the IDNA protocol itself. In this way, the LGRs can be used to validate all constraints on the label in one pass.

Common rules:

- Hyphen Restrictions — restricts the allowable placement of U+002D (-) HYPHEN (no leading/ending hyphen and no hyphen in 3-4 position). These constraints are described in section 4.2.3.1 of [RFC5891].

- Leading Combining Marks — restricts the allowable placement for combining marks (no leading combining mark). This constraint is described in section 4.2.3.2 of [RFC5891].

Rules for Right-to-Left labels:

- Leading Digit — restricts the allowable placement of digits for right-to-left labels (no leading digit in RTL label). This constraint is described in section 2.1 of [RFC5893].
- Mixed Digits — prevents the mixing of European and Arabic (Indic) digits. This constraint is described in appendix A.8 and A.9 of [RFC58932].

Context rules:

- Japanese in Label — constrains the code point at least one code point in the label containing such Katakana middle dot is from any of the Han, Hiragana, and Katakana scripts; This rule is described in Appendix A.7 of [RFC5892].

For these reference LGRs, the protocol-derived rules, other than the common rules, have only been included if they are needed for labels in the given script. The description section of each LGR file lists the rules and their associated references.

If a label to be validated has already been tested against protocol-derived constraints by the time the LGR is applied, these rules would be redundant and could be removed.

A small number of LGRs contain additional, LGR specific rules. These are documented in detail in the description section of the respective LGR.

Special rules:

- Extended-cp and Excluded-cp — as written, these two context rules always fail. That means, as written, the LGRs do not allow the code points identified as extended or excluded. Simply changing the appropriate rule so it always matches would enable the respective set code points without the need to edit the list of characters. Alternatively, the context condition could be removed from individual code points, thus enabling them one by one.

The purpose of these rules is thus to make the reference LGRs more adaptable to specific needs and to aid in review of the repertoire. See the description of extended and excluded code points above.

2.7 Metadata

The XML file format defines a number of elements for metadata. Several elements are not relevant to reference LGRs, but would be relevant to actual, deployed LGRs. These elements include <scope>, <validity-start>, and <validity-end>. For more details see [XML-LGR]. In adopting a reference LGR as the LGR for a specific zone, values for these elements should be supplied.

3 Expert Review

The LGRs were reviewed by independent reviewers with expertise in Unicode and linguistics, as well as IDNA and DNS security. The LGRs were updated to reflect the input from the review. These review reports are available at available at https://www.icann.org/resources/pages/second-level-lgr-2015-06-21-en.

4 Contributors

The reference LGRs were developed by the Staff and Contractors of Sheypa LLC.

1. Developers

Asmus Freytag
Michel Suignard

2. Expert Reviewers

Michael Everson
Nicholas Ostler
Lu Qin
Wil Tan

3. Community Members

Sheypa, LLC gratefully acknowledges the information provided by the following members of the community: TBD

5 References

[CLDR] CLDR - Unicode Common Locale Data Repository: http://cldr.unicode.org

[Guidelines] Internet Corporation for Assigned Names and Numbers, “Guidelines for Developing Reference LGRs for the Second Level”. (Los Angeles, California: ICANN, October 2015) https://www.icann.org/en/system/files/files/lgr-guidelines-second-level-30oct15-en.pdf.

[IIS] IIS, IDN Reference Tables, https://github.com/dotse/IDN-ref-tables

[XML-LGR] Davies, J and Asmus Freytag: “Representing Label Generation Rulesets using XML” https://www.ietf.org/id/draft-ietf-lager-specification-06.txt