Integration Panel:

Root Zone Label Generation Rules — LGR-2 Overview and Summary

REVISION – JULY 26, 2017

Table of Contents

1 Overview

This document describes the Label Generation Rules for the Root Zone (LGR) developed according to the "Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels" [Procedure]. The Procedure defines a two-stage process, in which community-based Generation Panels (GP) propose LGRs specific to a given script, which are then reviewed and integrated by the Integration Panel (IP). The result of the current round of this development work is the second version of the LGR (LGR-2), which is fully backwards compatible with [LGR-1].

The reader of this document is assumed to be familiar with the [Procedure][1], particularly the parts that describe the role of the IP and the tasks and expectations on the GPs.

The full content of LGR-2 is specified in a set of files as described in the next section.

1.1   Root Zone Label Generation Rules (LGR-2) Files

LGR-2 is provided as a collection of files that are self-contained and supersede the files from previous versions. The current document (https://www.icann.org/sites/default/files/lgr/lgr-2-overview-26jul17-en.pdf) provides background on the content and development of this version of the LGR. It also provides additional guidance to potential users of the LGR.

The normative definition of LGR-2 is provided as a set of XML files, consisting of one merged file and one XML file per script, as shown in Table 1.

Table 1. Merged (Common) and Element LGR files [XML – normative]

Script	File URL
Common	https://www.icann.org/sites/default/files/lgr/lgr-2-common-26jul17-en.xml
Arabic	https://www.icann.org/sites/default/files/lgr/lgr-2-arabic-script-26jul17-en.xml
Ethiopic	https://www.icann.org/sites/default/files/lgr/lgr-2-ethiopic-script-26jul17-en.xml
Georgian	https://www.icann.org/sites/default/files/lgr/lgr-2-georgian-script-26jul17-en.xml
Khmer	https://www.icann.org/sites/default/files/lgr/lgr-2-khmer-script-26jul17-en.xml
Lao	https://www.icann.org/sites/default/files/lgr/lgr-2-lao-script-26jul17-en.xml
Thai	https://www.icann.org/sites/default/files/lgr/lgr-2-thai-script-26jul17-en.xml

The Label Generation rules are expressed using a standard format defined in "Representing Label Generation Rulesets in XML" [RFC7940]. The remainder of this document assumes that the reader is at least familiar with some of the general concepts presented in that RFC.

The common LGR consists of a list of code points or sequences defining the merged repertoire as well as a set of mappings providing the variant relations between these repertoire items. In addition, the file contains a merged set of Whole Label Evaluation (WLE) rules for the root zone. Each code point in the file is annotated with the Unicode version in which it was first assigned, and the scripts in which it is used.

---

[1] References to documents cited are provided at the end.

Each of the script-specific files contains all the Label Generation Rules applicable to labels from that script, and only those rules. These files are called Element LGRs. Each file contains a description, a repertoire with optional variants, and WLE Rules, as well as detailed references that link each included code point to a reference that provides data justifying that code point's inclusion.

Table 2. Merged (Common) and Element LGR files [HTML - non normative]

Script	File URL
Common	https://www.icann.org/sites/default/files/lgr/lgr-2-common-26jul17-en.html
Arabic	https://www.icann.org/sites/default/files/lgr/lgr-2-arabic-script-26jul17-en.html
Ethiopic	https://www.icann.org/sites/default/files/lgr/lgr-2-ethiopic-script-26jul17-en.html
Georgian	https://www.icann.org/sites/default/files/lgr/lgr-2-georgian-script-26jul17-en.html
Khmer	https://www.icann.org/sites/default/files/lgr/lgr-2-khmer-script-26jul17-en.html
Lao	https://www.icann.org/sites/default/files/lgr/lgr-2-lao-script-26jul17-en.html
Thai	https://www.icann.org/sites/default/files/lgr/lgr-2-thai-script-26jul17-en.html

For each XML file, a mechanically generated and non-normative HTML presentation, as shown in Table 2, is provided for ease of review. This presentation is augmented by summary data, as well as data extracted from the Unicode Character Database [UCD], such as the character name.

Table 3. Other Files [PDF - non-normative]

Contents	File URL
Overview and Summary	This document
Repertoire Tables, non-CJK	https://www.icann.org/sites/default/files/lgr/lgr-2-non-cjk-26jul17-en.pdf

Repertoire tables are presented as non-normative PDF files that show the code points included in the repertoire for the merged LGR presented in the form of marked up tables. The presentation is similar to that used for character code charts in the Unicode Standard. The background color indicates the status of the code point:

Figure 1. Sample Repertoire Table

- Green: code points that are part of the LGR, including all members of code point sequences.[2]
- White: code points that are not PVALID in IDNA 2008 [RFC5892][IDNAREG].
- Pink: code points that are *excluded* from the Root Zone in a generic fashion (digits, hyphen), or by being excluded from the Maximal Starting Repertoire [MSR-2].
- Lavender: code points not included in the LGR as result of decisions by the Generation Panels during the development of the LGR.

Unicode blocks that contain no repertoire of the LGR are suppressed.

2   Process of Integration

2.1   Overview

The process for developing the Root Zone LGR consists of two stages, whereby a series of community-based Generation Panels creates and submits for public review a set of Proposed LGRs for their respective scripts. A separate expert panel, the Integration Panel, has the task to select from the submitted LGRs those ready for integration and to assemble them into a version of the Root Zone LGRs.

The [Procedure] assumes that each Generation Panel is best situated to make the selection of code points and variants specific to its script and to propose a disposition for them in the proposed LGR. In general, it is expected that Generation Panels will propose to include only a subset of code points that are in scope for their respective scripts. Generation Panels are expected to provide an adequate rationale including references for each code point included. See also [Guidelines].

The Integration Panel is tasked to evaluate the submitted in light of the Principles laid out in the [Procedure].

The review of LGR proposals undertaken by the Integration panel combines mechanical review steps with qualitative review in light of a set of principles as described in Section B.4 in [Procedure].

Mechanical review steps include verifying that the proposed LGR

- is within the MSR
- is within the scope (script)
- is symmetric and transitive (with respect to variants)
- contains all default WLE rules and actions
- contains the required files
- meets the syntax requirements

The qualitative review includes evaluation of the proposed LGR against these principles set out in Section A.3.6 in [Procedure] and [IABCP].

---

[2] If an LGR defines a code point sequence, but not the code point by itself, a valid label may contain that code point only as part of the given sequence. However, for the purpose of these repertoire tables, such code points are shown without any distinction.

Least Astonishment Principle: A Code Point in the Zone Repertoire should not present recognition difficulties to the zone's intended user population and should not lend itself to malicious use.

Contextual Safety Principle: A code point in the Zone Repertoire or any of its Variants that present unacceptable risks of being used in malicious ways should not be permitted.

Simplicity Principle: Overly complex rules are to be avoided, in favor of rules easily understood by users with only some background. In particular, in the root, rules should not require deep familiarity with a particular script or language.

Predictability Principle: People with reasonable knowledge of the topic should by and large reach the same conclusions about which code points should be included.

Stability Principle: Once a code point is permitted, it is almost impossible to stop permitting it: the act of permitting a code point cannot be undone. This is particularly true once a label containing this code point has been registered.

The following principles are normally satisfied implicitly, whether by the way the overall process is organized (by inclusion) or by the way the [MSR-2] defined the boundaries for LGRs. For the inclusion principle, in particular, the IP review checks whether all included code points are justified individually or by being part of a fixed set and documented as such.

Inclusion Principle: The zone repertoire is built up by specific inclusion; the default status for any code point is that it is excluded.

Letter Principle: Only Assigned Code Points normally used to write words should be permitted. Assigned Code Points normally used for both words and other purposes should not be permitted.

Longevity Principle: A Code Point in the Zone Repertoire should have stable properties across multiple versions of Unicode.[3]

The final principle is an overarching one that applies not only to code points, but also variants and other features of the LGR, and finally to the entire review and integration process. If there are doubts, it is best to withhold approval, rejecting or deferring a proposal until the doubt can be removed. From the Conservatism Principle also follows the prescription in [Procedure] to minimize allocatable variants and to maximize (within reason) the blocked variants.

Conservatism Principle: Any doubt should be resolved in favor of exclusion of a code point rather than inclusion.

For more details on the review carried out for specific proposals, see Section 2.3.

---

[3] Generally, that implies that code points from more recent versions of Unicode may require more stringent justification for inclusion.

2.2   Proposals Submitted

An integrated LGR starts from proposals for script-based LGRs. At the outset of the work on the current version of the Root Zone LGR the following proposals had been submitted by the respective Generation Panels:

Table 5. Script-Based LGR Proposals for the Root Zone

Script	Status	Files Submitted
Arabic	*in LGR-1*	arabic-lgr-proposal-18nov15-en.pdf
LGR Specification		proposed-arabic-lgr-18nov15-en.xml
Test Labels		arabic-labels-18nov15-en.txt
Armenian	*deferred*	armenian-lgr-proposal-05nov15-en.pdf
LGR Specification		proposed-armenian-lgr-05nov15-en.xml
Test Labels		armenian-test-labels-05nov15-en.txt
Ethiopic	*accepted*	proposal-ethiopic-lgr-17may17-en.pdf
LGR Specification		proposal-ethiopic-lgr-17may17-en.xml
Test Labels		ethiopic-test-labels-17may17-en.txt
Georgian	*accepted*	proposal-georgian-lgr-24nov16-en.pdf
LGR Specification		proposal-georgian-lgr-15sep16-en.xml
Test Labels		georgian-test-labels-15sep16-en.txt
Khmer	*accepted*	proposal-khmer-lgr-15aug16-en.pdf
LGR Specification		proposal-khmer-lgr-15aug16-en.xml
Test Labels		khmer-test-labels-15aug16-en.txt
Lao	*accepted*	proposal-lao-lgr-31jan17-en.pdf
LGR Specification		proposal-lao-lgr-31jan17-en.xml
Test Labels		lao-test-labels-31jan17-en.txt
Thai	*accepted*	proposal-thai-lgr-25may17-en.pdf
LGR Specification		proposal-thai-lgr-25may17-en.xml
Test Labels		thai-test-labels-25may17-en.txt

The Integration Panel reviewed these proposals and determined whether they could be integrated into the current version of the LGR.

2.3   Review of Proposals

2.3.1   General Notes on the Proposal Review

After a thorough review, the Integration Panel was unanimous in accepting the LGRs for Ethiopic, Georgian, Khmer, Lao and Thai for integration into LGR-2. The Integration Panel unanimously continued the deferral of the LGR for Armenian, because its interaction with other scripts that cannot be evaluated at this time. That proposal is not rejected, but deferred for review in the context of a future LGR.

The Arabic LGR had been reviewed and approved for integration into LGR-1. It is retained for LGR-2.

As result of the review of proposals submitted, the contents of LGR-2 are defined by six script-specific LGRs listed as accepted or retained from LGR-1 in Table 5 above, in addition to the default WLE rules and actions defined by the Integration Panel (IP) as part of the [MSR-2]. (See Section 3 for a summary of the contents of the Root Zone LGR).

See the following subsections for details on the review and disposition of specific proposals for each script. Please note:

(a) Details on the review of proposals from any previous edition of the LGR are not repeated here.
(b) The summary of the reviews of scripts included for the first time in this edition of the LGR each cover the following points:
    - Overview,
    - Highlight of particular issues encountered,
    - Scope of mechanical testing of LGR proposal,
    - Scope of label testing,
    - Potential for collisions with code points in any other script, and
    - Disposition.

2.3.2   Arabic LGR Proposal
For information on the original review of [Proposal-Arabic], see Section 2.3.2 of [LGR-1].

The Arabic Script LGR has been part of the Root Zone LGR since [LGR-1]. Being upwardly compatible, the current version continues to include this script LGR unchanged from [LGR-1], except for minor editorial adjustments.

2.3.3   Armenian LGR Proposal
For information on the original review of [Proposal-Armenian], see Section 2.3.1 of [LGR-1].

While the Armenian LGR proposal was successfully submitted and passed mechanical and other review, the IP continues in the conclusion, that the script should be treated as being related to other scripts in the sense of Section 3.2 of MSR-2. Consequently, the IP chose to continue to defer the script until its interactions with the related scripts are well-enough understood to cause no risk of future incompatibilities.

2.3.4   Ethiopic LGR Proposal
The Integration Panel worked with the Ethiopic Generation Panel [EthiopicGP] during the development of [Proposal-Ethiopic] to ensure that it would meet the Integration Panel's understanding of the [IABCP] principles and other prescriptions found in [Procedure].

In particular, the review focused on the definition of languages to be supported (with the Ethiopic script); the recognition of homophones in Amharic as variants; the impact of variants designed specifically for one language on other languages using the script; and determining the relevance of certain putative candidates for cross-script homoglyphs.

A separate mechanical review of the proposal has verified that the specification of the repertoire in the XML is valid and in accordance with [Proposal-Ethiopic]; it has further confirmed, by evaluating the supplied test labels, that the result of applying the LGR adequately reflects the understanding that went into its design.

The LGR was also reviewed against a set of putative Ethiopic labels derived from corpora in two languages (Amharic and Tigrigna), as well as against any existing Ethiopic ccTLDs and gTLDs. This review included some verification of any wider impact on languages besides Amharic of the blocking variants introduced. These variants were introduced to accommodate vagueness in the spelling of homophones in the Amharic language.

The Ethiopic script was judged a separable script, according to section 3.3 of [MSR-2]; it can be integrated in isolation, without the risk of introducing future incompatibilities.

Based on this review and having resolved any open issues in discussion with the Ethiopic-GP, the IP unanimously decided that the Ethiopic LGR Proposal is ready for integration into the Root Zone LGR as submitted

2.3.5   Georgian LGR Proposal

The Integration Panel worked with the Georgian Generation Panel [GeorgianGP] during the development of [Proposal-Georgian] to ensure that it would meet the Integration Panel's understanding of the [IABCP] principles and other prescriptions found in [Procedure].

In particular, this included reviewing the question whether any of the minority languages required additions to the repertoire.

A separate mechanical review of the proposal has verified that the specification of the repertoire in the XML is valid and in accordance with [Proposal-Georgian]; it has further confirmed, by evaluating the supplied test labels, that the result of applying the LGR adequately reflects the understanding that went into its design.

The LGR was also reviewed against any existing Georgian ccTLDs and gTLDs

The Georgian script was judged a separable script, according to section 3.3 of [MSR-2]; it can be integrated in isolation, without the risk of introducing future incompatibilities.

Based on this review and having resolved any open issues in discussion with the Georgian GP, the IP unanimously decided that the Georgian LGR Proposal is ready for integration into the Root Zone LGR as submitted

2.3.6   Khmer LGR Proposal

The Integration Panel worked with the Khmer Generation Panel [KhmerGP] during the development of [Proposal-Khmer] to ensure that it would meet the Integration Panel's understanding of the [IABCP] principles and other prescriptions found in [Procedure].

In particular, the review focused on the criteria for excluding code points on grounds of restricted use; the extent of WLE rules needed to express ordering constraints on vowels and diacritics within a syllable (specifically to avoid coding ambiguity hidden by inexact rendering, and to avert rendering failures). Further attention was paid to ways of simplifying the specification of context rules in light of the Simplicity and Predictability principles. Another issue was the relative merits of defining code point sequences as instead of WLE rules; and in defining variant relations between sequences of characters. The Khmer script was considered in the context of both Thai and Lao scripts, to reduce the risk of introducing incompatibilities.

A separate mechanical review of the proposal has verified that the specification of the repertoire and WLE rules in the XML are valid and in accordance with [Proposal-Khmer]; it has further confirmed, by evaluating the supplied test labels, that the result of applying the LGR adequately reflects the understanding that went into its design.

The LGR was also reviewed against a set of putative Khmer labels derived from a text corpus, as well as against any existing Khmer ccTLDs and gTLDs.

Based on this review and having resolved any open issues in discussion with the Khmer GP, the IP unanimously decided that the Khmer LGR Proposal is ready for integration into the Root Zone LGR as submitted

2.3.7   Lao LGR Proposal

The Integration Panel worked with the Lao Generation Panel [LaoGP] during the development of [Proposal-Lao] to ensure that it would meet the Integration Panel's understanding of the [IABCP] principles and other prescriptions found in [Procedure].

In particular, the review focused on the appropriate representation of constraints on sequences of vowel code points (specifically to avoid impossible or ambiguous combinations, and to avert rendering failures). This included developing a distinction between required structural constraints as opposed to orthographic restrictions (spelling rules), and the best ways to simplify the specification of context rules in light of the Simplicity and Predictability principles. The Lao script was considered in the context of both Khmer and Thai scripts, to reduce the risk of introducing incompatibilities.

A separate mechanical review of the proposal has verified that the specification of the repertoire and WLE rules in the XML are valid and in accordance with [Proposal-Lao]; it has further confirmed, by evaluating the supplied test labels, that the result of applying the LGR adequately reflects the understanding that went into its design.

The LGR was also reviewed against a set of putative Lao labels derived from a text corpus, as well as against any existing Lao ccTLDs and gTLDs.

Based on this review and having resolved any open issues in discussion with the Lao GP, the IP unanimously decided that the Lao LGR Proposal is ready for integration into the Root Zone LGR as submitted.

2.3.8   Thai LGR Proposal

The Integration Panel worked with the Thai Generation Panel [ThaiGP] during the development of [Proposal-Thai] to ensure that it would meet the Integration Panel's understanding of the [IABCP] principles and other prescriptions found in [Procedure].

In particular, the review focused on whether minority languages when written with the Thai script require adjustments to the context rules and the appropriateness of the representation of constraints on sequences of diacritic code points (specifically to avoid coding ambiguity cloaked by rendering, and to avoid rendering failures). This included the proper balance between the desires to be inclusive, by fully supporting other languages using the Thai script and to address security by imposing restrictions necessary to ensure stable rendering. The Thai script was considered in the context of both Khmer and Lao scripts, to reduce the risk of introducing incompatibilities.

A separate mechanical review of the proposal has verified that the specification of the repertoire and WLE rules in the XML are valid and in accordance with [Proposal-Thai]; it has further confirmed, by evaluating the supplied test labels, that the result of applying the LGR adequately reflects the understanding that went into its design.

The LGR was also reviewed against a set of putative Thai labels derived from a text corpus, as well as against any existing Thai ccTLDs and gTLDs.

Based on this review and having resolved any open issues in discussion with the Thai GP, the IP unanimously decided that the Thai LGR Proposal is ready for integration into the Root Zone LGR as submitted.

3   Integration and Contents of LGR-2

3.1   General Notes

After reviewing and accepting a proposed LGR, the Integration panel prepares an XML file containing an equivalent LGR as measured in terms of valid labels and variants produced, but with changes to the metadata and comments for consistency with the other elements of an integration process for the Root Zone LGRs. Collectively, these constitute the Element LGRs. From each an annotated HTML file is created mechanically for a more human-readable presentation of the data.

Element LGRs included from earlier versions of the LGR are updated as to version number and date; minor changes to other metadata and comments for consistency are also applied.

From the Element LGRs a merged XML file is created mechanically containing the union of the repertoire and non-reflexive variant mappings and annotating each item in the repertoire and rules to mark its origin in a particular element LGR. This file constitutes the Common LGR. Because the actual type of all variant mappings is script-specific and therefore cannot be represented in a merged file, all variant mappings are set to "blocked" in the merged file (See also Section 5).

While script-specific tags, rules and classes are prefixed with a script name and individually included, all actions and default WLE rules from the Element LGRs are coalesced in the merged file. In principle, the default WLE rules and any actions are not script-specific, but in practice, they are usually triggered by ranges of code points or variant types specific to an element LGR. The IP manually reviews the result to make sure that these elements from different LGRs do not conflict. If necessary, they are restated. Finally, an annotated, human-readable presentation of the merged file is created.

The following sections summarize briefly the contents of particular files making up the Root Zone LGR. These files are listed in Tables 1 and 2 above. For more details and background on the organization of the LGR across files, see [Packaging].

3.2   Merged LGR (Common)

3.2.1   Repertoire

The repertoire of the merged Root Zone LGR is the cumulative repertoire of all the Element LGRs that have been integrated into this version. Those repertoires, in turn were developed based on [MSR-2], which is a subset of Unicode 6.3 [Unicode 6.3], that excludes code points used for historical or special purposes only, or those used in languages that did not meet the criteria for stable and modern usage as outlined in [MSR-2].

As appropriate for the Root Zone LGR, the repertoire includes neither digits nor the HYPHEN-MINUS.

The merged repertoire contains all sequences defined by the Element LGRs. If any code point that is a member of a sequence is not also listed by itself in an Element LGR, it will not be defined by itself in the merged LGR. Root Zone labels may contain that code point, but only as part of a defined sequence.

3.2.2   Variants

The variant mappings in the merged LGR are the union of the non-reflexive variant mappings from all the Element LGRs that have been integrated into this version of the Root Zone LGR. Because the disposition of variant labels, for example as "allocatable", is specific to each script, they cannot be expressed in the script-neutral context of this integrated LGR. Instead, in the merged LGR, all variant mappings are given the type "blocked". (This allows the use of the Common LGR in checking for conflicts between labels as described in Section 5.4.)

3.2.3   Character Classes

The character classes in the merged LGR are the union of the character classes from all the Element LGRs that have been integrated into this version of the Root Zone LGR. Many character classes are derived in turn from tag values associated with code points in the repertoire. These tag values have also been merged. To avoid duplications, the names of all tags and character classes in the merged LGR are prefixed by the four-letter Unicode script identifier identifying the Element LGR from which they were merged.

3.2.4   Whole Label Evaluations (WLE) Rules

The merged LGR includes the cumulative set of Whole Label Evaluation rules and actions for all Element LGRs that have been integrated into this version. WLE rules include both context rules and whole-label rules. The purpose of WLE rules and actions for the Root Zone LGR is to allow automatic exclusion of labels that present particular challenges in display and processing, such as a label leading off with a combining mark, because that mark would tend to combine visually with the code point in front of it. Based on [Procedure] the Root Zone LGR has a single set of WLE rules that is common to all scripts. In practice, most rules are written to be specific to only the code points encountered in labels of a given script, so that the rules do not interact with each other. Each Element LGR only contains rules that are specified to it (as well as any default rules) while the IP has reviewed and made sure that the combined rules in the merged LGR do not give rise to conflicts.

To make the merged set of rules easier to follow and to avoid unintentional naming conflicts, the names of any context or whole label rules defined by an Element LGR have been prefixed by the four-letter Unicode script identifier for that LGR before being merged into the Common LGR. The same has been done for tags and character classes. Finally, all repertoire code points have been tagged with their script extension (scx) values[4] before being merged.

[MSR-2] defines a number of default rules and actions. These are present in all Element LGRs and in the Common LGR. They have been annotated in the Common LGR with the prefix "Common-".

Actions are merged, preserving their relative order of precedence from the Element LGR.

For additional details on the merged LGR, see Section 5.3 below.

The following subsections give a brief summary of the contents of each of the Element LGRs contained in this version of the Root Zone LGR. The full definition of the element LGRs is provided in files listed in Tables 1 and 2 above. (In addition, the files in Table 3 above provide a visual summary of the contents of the repertoire of the Root Zone LGR).

3.3   Arabic Element LGR

3.3.1   Repertoire for Arabic

The repertoire for the Arabic Element LGR is described in Section 3.2 in [Proposal-Arabic] by the Task Force for Arabic IDNs [TF-AIDN]. It includes only the 128 code points used by languages that are actively written in the Arabic script. It excludes code points for which TF-AIDN was unable to find sufficient evidence of use (see Appendix F in [Proposal-Arabic]).

The Arabic Element LGR does not include combining marks or code point sequences. All combining marks have been excluded for these reasons:

---

[4] Code points used with more than one script have "scx" values consisting of a list of script identifiers. For the Root Zone LGR, script identifiers not associated with the Root Zone are suppressed.(See [UCD])

- First, they can significantly overproduce and would require additional rules to constrain them effectively, complicating the design.
- Second, even where they are required for some languages, they are optional for others.
- Third, this also circumvents the issue regarding duplication between some precomposed code points and combining sequences raised by [IAB].

As part of the Root Zone, the element LGR includes neither digits nor the HYPHEN-MINUS.

For further details, see Section 3.2 "Code point repertoire included", in [Proposal-Arabic].

The Arabic LGR was first included in [LGR-1].

3.3.2   Variants for Arabic

The Arabic Element LGR includes "blocked" and "allocatable" variants, assigned according to Section 4 "Final recommendation of variants for Top Level Domains (TLDs)" in [Proposal-Arabic]. These recommendations balance the desire to minimize the number of possible allocatable variants with the need to keep the definition of variants simple.

3.3.3   Whole Label Evaluation Rules for Arabic

The Arabic Element LGR includes Whole Label Evaluation rules specific to the Arabic script. See Section 5 "Whole Label Evaluation (WLE) rules", in [Proposal-Arabic]. As specified, these rules serve to prevent the mixing of two variants of the same code point within the same label. This has the effect of reducing overproduction of variant labels. See also the comments given for each rule or action.

3.3.4   Default Whole Label Evaluation Rules

The Arabic Element LGR includes the set of required default WLE rules and actions applicable to the Root Zone and defined in [MSR-2].

3.4   Ethiopic Element LGR

3.4.1   Repertoire for Ethiopic

The repertoire for the Ethiopic Element LGR is described in Section 5 of [Proposal-Ethiopic]. It includes only the 311 code points from the Ethiopic script needed to write languages commonly using the Ethiopic script.

The element LGR does not include combining marks or sequences.

As part of the Root Zone, the element LGR includes neither digits nor the HYPHEN-MINUS.

3.4.2   Variants for Ethiopic

As described in Section 6 of [Proposal-Ethiopic], the element LGR includes a number of variants for code points that are homophones in Amharic.

3.4.3   Whole Label Evaluation Rules for Ethiopic

The element LGR includes no script-specific WLE rules.

3.4.4   Default Whole Label Evaluation Rules

The Ethiopic Element LGR includes the set of required default WLE rules and actions applicable to the Root Zone and defined in [MSR-2].

3.5   Georgian Element LGR

3.5.1   Repertoire for Georgian

The repertoire for the Georgian Element LGR is described in Section 5 of [Proposal-Georgian]. It includes only the 33 code points from the Mkhedruli alphabet that are needed to write modern Georgian, a set also sufficient to write the other languages widely used and commonly written with the Georgian script.

The element LGR does not include combining marks or sequences.

As part of the Root Zone, the element LGR includes neither digits nor the HYPHEN-MINUS.

3.5.2   Variants for Georgian

The element LGR includes no variants.

3.5.3   Whole Label Evaluation Rules for Georgian

The element LGR includes no script-specific WLE rules.

3.5.4   Default Whole Label Evaluation Rules

The Georgian Element LGR includes the set of required default WLE rules and actions applicable to the Root Zone and defined in [MSR-2].

3.6   Khmer Element LGR

3.6.1   Repertoire for Khmer

The repertoire for the Khmer Element LGR is described in Section 5 of [Proposal-Khmer]. It includes only the 71 code points used to write modern languages in widespread common use and commonly written in the Khmer script.

The Khmer script is a complex script using consonants as base letters and combining marks for vowels and other signs. A special combining mark, U+17D2 KHMER SIGN COENG, forms sequences with following consonants that are to be rendered as subscripted form. The Khmer Repertoire explicitly lists two of these subjoined consonant sequences because of the variant relationship established between them.

As part of the Root Zone, the element LGR includes neither digits nor the HYPHEN-MINUS.

3.6.2   Variants for Khmer

The Khmer Element LGR includes two sequences for subjoined consonants that are "blocked" variants of each other due to identical appearance. When not subjoined, these consonants are not variants of each other. See Section 6 in [Proposal-Khmer].

3.6.3   Whole Label Evaluation Rules for Khmer

The Khmer script uses combining marks for vowels and other signs. These signs cannot occur in all contexts and the Khmer Element LGR uses context rules defined in Section 7 of [Proposal-Khmer] to prevent their occurrence in contexts that could give rise to security risks; also defined is a whole label rule limiting the number of adjacent subjoined consonant sequences.

3.6.4   Default Whole Label Evaluation Rules

The Khmer Element LGR includes the set of required default WLE rules and actions applicable to the Root Zone and defined in [MSR-2].

3.7   Lao Element LGR

3.7.1   Repertoire for Lao

The repertoire for the Lao Element LGR is described in Section 5 of [Proposal-Lao]. It includes only the 51 code points used to write modern languages in widespread common use and commonly written in the Lao script.

The Lao script is a complex script using consonants as base letters and combining marks for vowels and other signs. The Lao Repertoire explicitly lists one sequence of vowel marks because it occurs in a specific context.

As part of the Root Zone, the element LGR includes neither digits nor the HYPHEN-MINUS.

3.7.2   Variants for Lao

The element LGR includes no variants.

3.7.3   Whole Label Evaluations Rules for Lao

The Lao script uses combining marks for vowels, tone marks and other signs. These signs cannot occur in all contexts and the Lao Element LGR uses context rules defined in Section 7 of [Proposal-Lao] to prevent their occurrence in contexts that could give rise to security risks; also defined is a context rule limiting the number of adjacent repetition marks at the end of the label.

To reduce complexity, the rules allow many labels that users would reject as impossible to occur in the context of writing Lao, but that represent no security risk. In contrast, a small number of words cannot be represented as labels under this LGR; a tradeoff deemed acceptable to the Lao GP as accommodating them would have required special cases to be added to the rules.

3.7.4   Default Whole Label Evaluation Rules

The Lao Element LGR includes the set of required default WLE rules and actions applicable to the Root Zone and defined in [MSR-2].

3.8   Thai Element LGR

3.8.1   Repertoire for Thai

The repertoire for the Thai Element LGR is defined in Section 5 of [Proposal-Thai]. It includes only the 69 code points used to write modern languages in widespread common use and commonly written in the Thai script.

The Thai script is a complex script using consonants as base letters and combining marks for vowels and other signs. The Thai Repertoire explicitly lists one sequence of vowel marks and two sequences of consonants because they occur in a specific context. One code point, U+0E45, only occurs as part of a sequence; thus, it is not listed by itself as a member of the repertoire.

The code point U+0E33, representing one of the Thai vowels, is DISALLOWED in IDNA 2008. In labels, this code point must be expressed as the sequence U+0E30 U+0E4D instead. This sequence is explicitly a member of the repertoire, to allow the exceptional occurrence of U+0E4D after a specific above-vowel.

As part of the Root Zone, the element LGR includes neither digits nor the HYPHEN-MINUS.

3.8.2   Variants for Thai

The element LGR includes no variants.

3.8.3   Whole Label Evaluations Rules for Thai

Thai is a complex script which a set of code points create a character-cluster in a cell, and only a subset of all possible code point sequences would ever expected to occur. However, the WLE rules defined in Section 7 of [Proposal-Thai] are used to limit the contexts in which certain code points (including some consonants, vowels, tone and diacritics) may appear in the coded sequence. These ensure that the characters occur in the order expected (and supported) by typical rendering engines: they are not intended to enforce 'spelling-rules'.

The whole label evaluation rules for Thai would need to be relaxed over those in use for the Thai language to fully cover patterns that occur in other languages using the Thai script. However, that is not possible due the fact that unstable rendering for those patterns creates a security concern, where rendering presently becomes unreliable.

To use the simple generalized WLE Rules will also allow the user of other languages to be able to input a string in their language using the Thai Script without any limitation like spelling rules, while maintaining the consistent ordering expected by rendering engines.

3.8.4   Default Whole Label Evaluation Rules

The Thai Element LGR includes the set of required default WLE rules and actions applicable to the Root Zone and defined in [MSR-2].

4 General Notes on the Root Zone LGR

4.1 Rules

"Label Generation Rules (LGR)" is the term used to describe the sets of code points, and the constraints on them, that are needed to generate IDNs in a particular script (e.g. Latin, Arabic, or Japanese).

Most of the information in a typical LGR takes the form of selections from a repertoire of code points defined in the Unicode Standard, further reduced by [MSR-2] in the case of the Root Zone. The "R" in LGR stands for "Rules" rather than "Repertoires", because labels must be constructed out of permitted code points in context, including allowing sequences of code points as repertoire items. The validity of labels is determined by mechanically evaluating the LGR, and in particular, the Whole Label Evaluation (WLE) rules, which use the wider context of a label. In addition, variant rules define what variant labels might exist and whether they are or are not available for allocation.

4.2 Scripts

In defining labels fit to be used globally in the DNS root zone, any code point is defined as belonging to a script, with some code points used with multiple scripts, as defined by the Script_Extensions property in the Unicode Character Database [UCD]. For the root zone, all code points used in a given label must normally belong to a single script; although any script supported in the LGR may be used to create a root label, and those labels can in principle be used anywhere in the internet, there cannot be a mixture of scripts represented within a single root label. Notably, for example, root zone LGRs for any script other than Latin cannot introduce US-ASCII code points into their repertoire.

The definition of script for used in the LGR process is that chosen by [ISO 15924]; this definition recognizes that Japanese and Korean are written with a mixture of scripts, respectively a mixture of Han ideographs with Kana or Hangul, and provides separate script identifiers for such composite scripts.

Many scripts, such as Arabic, Cyrillic, Devanagari and Latin each support a variety of languages. As long as the code points are members of the same script, as defined by [ISO15924], code points used for different languages can be mixed in a label; subject only to constraint on mixing that might be present in the WLE rules of the respective LGR.

4.3 Comprehensiveness and Staging

Ideally, the Root Zone LGR would be comprehensive, that is, include all scripts eligible for the root zone from its first version. With respect to the *Stability Principle* and the *Least Astonishment Principle* [IABCP] a fully comprehensive LGR would guarantee that all issues relating to the possible interaction among all scripts can be fully investigated in the development of the LGR. From a practical perspective doing so would be prohibitive because of the additional time needed to investigate certain scripts, and perhaps unnecessary for two main reasons.

First, not all scripts are related closely enough so that they affect each other from the perspective of LGR development. Second, it is not realistic to expect that Generation Panels will be formed and complete

their work for all eligible scripts within the same time frame. Consequently, the [Procedure] anticipated that LGR would be rolled out in stages.

The goal for all future versions of the LGR must be to retain full backward compatibility, so that they preserve the output of any label registration against the old LGR, when applied to an updated LGR. Consequently, the IP anticipates that succeeding versions of the LGR will be strict supersets of its predecessors. It is expected that registrations that predate the initial release of an LGR covering the respective script will be allowed to remain, even if in conflict, but without becoming a binding precedent for the LGR itself. To date, there is no known instance of such a conflict.

5   Using the LGR

5.1   Element LGRs

The merged file containing the Common LGR and the per-script Element LGRs serve different purposes. At the time of registration, the applicant selects the script in the context of which the label is to be applied. That selection determines which element LGR is used in processing the application. Each script-specific element LGRs presents the complete data and specification to determine the validity of a label as well as to generate the full set of allocatable variants for the label, when applied for under that script.

5.2   Common LGR

The Common (merged) LGR contains the cumulative repertoire, WLE rules and all non-reflexive variant mappings (with type set to "blocked). The merged LGR thus presents the complete data and specification needed for conflict checking with any existing label in the Root Zone, independent of script.

Note that the merged LGR cannot be used to determine the validity of a label, because the validity of a label depends directly on the specific subset of the overall repertoire that is defined for a given script. (Simply applying the merged LGR would result in returning mixed script labels as valid). The validity of a label may further depend in some circumstances on the script-specific definition of variants. For these reasons, the merged LGR cannot be used for final validity checking of a label. (A label reported as invalid by the merged LGR is also invalid at the Element LGR level, but the reverse does not hold.)

5.3   Other uses of the Common LGR

As outlined above, the merged LGR serves mainly in the detection of collisions between applied for and delegated (or reserved) labels. In addition, the merged LGR provides:

- documentation of the overall repertoire; in addition to formal data definition in the XML file, and the annotated repertoire table in the HTML, the data from the merge are also used to drive the production of the PDF overview charts;
- documentation of the overall system of WLE rules and actions. The merged rule sets document that rules for different scripts are not in conflict with each other for the same code point;
- an index relating code points to script LGRs; as the "script" from an LGR perspective is not a true partition of the repertoire, particularly for CJK, the common LGR is the way to quickly look up which script LGRs support a code point;

- a starting point for getting from any supported code point in the Root Zone to the description in the various proposal documents and from there to the background documents on which inclusion of these code points is based. To this end, the "ref" attributes identify the relevant proposal for each code point, variant, class, rule and action.

5.4   Steps in Processing a Label

In order to determine the disposition of a label, it is evaluated against the Root Zone LGR in three steps.

1. *Verify that a proposed label is valid by processing it with the Element LGR corresponding to the script that was selected for the label in the application.*

   This check will determine whether all code points in the label are defined in the LGR, and whether each code point meets all the context rules defined for it. In addition, all whole-label rules are evaluated; if a disposition other than "valid" results, the label is invalid.

   At this first step, do not enumerate all variants. However, as part of checking validity it is necessary to evaluate any reflexive variants, and apply dispositions based on their types. For example, if any reflexive variant is of type "out-of-repertoire-var", the label will be invalid.

   For any invalid label, stop the processing.

2. *Process the now validated label against the common LGR to verify it does not collide with any existing delegated labels (and any of their variants, whether blocked or allocatable).*

   Each label and all its variants form a variant label set. For the Root Zone LGR, all variant label relations are symmetric and transitive, which means that all such variant label sets are disjoint (do not overlap). For each label, calculate an 'index' label identifying the set (for example the element lowest in code point order). Any two labels resulting in the same index label will collide: either with each other or with one of the variants of the other label.

   For any label that collides with existing labels, stop the processing.

3. *Now that the label is known to be valid, and not in collision, use the appropriate element LGR to generate all allocatable variants.*

   The valid label and any allocatable variants constitute the result of the LGR processing and form the input into any subsequent stages of the application and registration process.

5.5   Design Notes for the Root Zone LGR

5.6   Reducing Complexity

As much as possible, the variant mappings and types in the Element LGRs have been drawn up to limit the number of allocatable variants generated. Where applicable, WLE rules reduce the number of valid labels, and in some cases, they reduce the number of allocatable variants as well. Both mechanisms typically rely on dividing the variants by some suitable linguistic context and on mechanically preventing the mixing of variants from different contexts in the same label.

To reduce complexity of the ruleset, some loss in linguistic fidelity has been accepted where it resulted in simpler rules that do not compromise security. Where possible, constraints have been presented as context rules on code points or via enumeration of sequences in the repertoire. Where context rules are used, those implementing constraints on immediately following or preceding code points have been preferred: no attempt is being made, for example, to implement full segmentation into valid syllables.

In accordance with the [Procedure] the LGR is designed to mechanically eliminate (or take off the plate) as much as possible any labels and variant labels that pose an undue risk to the usability and security of the DNS.

5.7   Limitations of the LGR

There are limitations to what can be done with mechanical application of rules, and in some cases, it is not possible to reduce the number of allocatable labels that is practicable and safe without creating undue restrictions on otherwise valid labels. In this context, it is a useful reminder that having a label that is "allocatable" means neither that it will necessarily be delegated, nor that it necessarily should be delegated. In fact, investigations of actual registrations on the second level reveal that applicants have tended to apply for only a small number of variant labels.

The LGR can be thought of as creating a maximal set of valid labels and allocatable variants, but other steps in the registration process are expected to include suitable mechanisms to further reduce the list of labels available for delegation. It is the view of the Integration Panel that such reduction is necessary, because the larger the number of delegated variants the larger the risk they create to the DNS.

6 Summary of Changes

6.1   Changes by revision

1. LGR-1 added 128 code points for 1 script, plus 17 WLE rules and 21 actions.
2. LGR-2 added 535 code points for 5 scripts, plus 28 WLE rules and 1 action.

6.2   Code points by script

The following table shows how many code points, by script, are available for root zone LGR development by being included in [MSR-2] and how many are selected for each version of the LGR. The count includes code points that are only available as part of a defined sequence.

Script tag	Script Name	MSR-2	LGR-1	LGR-2
Arab	Arabic	239	128	128
Armn	Armenian	38		
Beng	Bengali	64		
Cyrl	Cyrillic	93		
Deva	Devanagari	91		
Ethi	Ethiopic	364		311
Geor	Georgian	37		33
Grek	Greek	36		
Gujr	Gujarati	66		
Guru	Gurumukhi	61		
Hang	Hangul	11172		
Hani	Han Ideographs	19852		
Hebr	Hebrew	46		
Hira	Hiragana	89		
Kana	Katakana	92		
Khmr	Khmer	78		71
Knda	Kannada	68		
Laoo	Lao	53		51
Latn	Latin	305		
Mlym	Malayalam	73		
Mymr	Myanmar	90		
Orya	Oriya	66		
Sinh	Sinhala	79		
Taml	Tamil	49		
Telu	Telugu	67		
Thaa	Thaana	50		
Thai	Thai	71		69
Tibt	Tibetan	80		
Zinh	INHERITED	21		
Total		33490	128	663

23

7 Contributors

LGR-2 and its precursor versions were developed by the Integration Panel, based on proposals submitted by the respective Generation Panels, with input from community members, as well as support by ICANN staff members. The following lists of contributors are cumulative.

7.1   Integration Panel Members

Marc Blanchet
Asmus Freytag
Nicholas Ostler
Michel Suignard
Wil Tan

7.2   Advisors

Lu Qin

7.3   Community Members

The integration panel gratefully acknowledges the information provided by the following members of the community:

Members of TF-AIDN (Arabic)
Members of the Armenian Generation Panel
Members of the Ethiopic Generation Panel
Members of the Georgian Generation Panel
Members of the Khmer Generation Panel
Members of the Lao Generation Panel
Members of the Thai Generation Panel
Olivier Crepin-Leblond
Chris Dillon
Yuriy Kargapolov
Narine Khachatryan
Meikal Mumin
Dusan Stojicevic
Richard Wordingham

7.4   ICANN Staff

Sarmad Hussain
Alireza Saleh
Anand Mishra
Jia-Juh Kimoto
Kim Davies

8   References

[Armenian GP] Armenian Script Generation Panel, see Section 8 of [Proposal-Armenian]

[Ethiopic GP] Ethiopic Script Generation Panel, see Section 8 of [Proposal-Ethiopic]

[Georgian GP] Georgian Script Generation Panel, see Section 8 of [Proposal-Georgian]

[Khmer GP] Khmer Generation Panel, see Section 8 of [Proposal-Khmer]

[Lao GP] Lao Generation Panel, see Section 8 of [Proposal-Lao]

[Thai GP] Thai Generation Pane, see Section 8 of [Proposal-Thai]

[Guidelines] Internet Corporation for Assigned Names and Numbers, "Guidelines for Developing Script-Specific Label Generation Rules for Integration into the Root Zone LGR". (Los Angeles, California: ICANN, December 2014) https://community.icann.org/download/attachments/43989034/Guidelines-for-LGR-2014-12-02.pdf.

[IAB] Internet Architecture Board (IAB), "IAB Statement on Identifiers and Unicode 7.0.0" https://www.iab.org/documents/correspondence-reports-documents/2015-2/iab-statement-on-identifiers-and-unicode-7-0-0/

[IABCP] Sullivan, A., *et al.,* "Principles for Unicode Code Point Inclusion in Labels in the DNS". Internet Architecture Board (IAB) = RFC 6912 http://tools.ietf.org/html/rfc6912.

[IDNAREG] IANA Registry: "IDNA Parameters". For Unicode 6.3 available at: http://www.iana.org/assignments/idna-tables-6.3.0/idna-tables-6.3.0.xml. Visited 2013-11-20.

[ISO15924] *Codes for the representation of names of scripts*, ISO 15924:2004. Available from http://www.unicode.org/iso15924/. Visited 2012-09-21.

[LGR-1] Integration Panel, "Integration Panel: Root Zone Label Generation Rules — LGR-1", 24 February 2016 https://www.icann.org/sites/default/files/lgr/lgr-1-overview-24feb16-en.pdf

[MSR-2] Integration Panel, "Maximal Starting Repertoire — MSR-2 Overview and Rationale", 14 April 2015 https://www.icann.org/en/system/files/files/msr-2-overview-14apr15-en.pdf

[Packaging] Integration Panel: "Packaging the MSR and LGR", 24 April 2015, https://community.icann.org/download/attachments/43989034/Packaging-MSR-LGR.pdf

[Procedure] Internet Corporation for Assigned Names and Numbers, "Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels." (Los Angeles, California: ICANN, March, 2013) http://www.icann.org/en/resources/idn/variant-tlds/draft-lgr-procedure-20mar13-en.pdf

[Proposal-Arabic] TF-AIDN, "Proposal for Arabic Script Root Zone LGR", Version 3.4, 18 November 2015, Supporting Document https://www.icann.org/en/system/files/files/arabic-lgr-proposal-18nov15-en.pdf [PDF, 3.47 MB]

[Proposal-Armenian] Armenian Generation Panel, "Proposal for Armenian Script Root Zone LGR", 05 November 2015 https://www.icann.org/en/system/files/files/armenian-lgr-proposal-05nov15-en.pdf

[Proposal-Ethiopic] Ethiopic Script Generation Panel, "Proposal for Ethiopic Script Root Zone LGR", 17 May 2017, Supporting Document: https://www.icann.org/en/system/files/files/proposal-ethiopic-lgr-17may17-en.pdf

[Proposal-Georgian] Georgian Script Generation Panel, "Proposal for the Georgian Script Root Zone LGR", 24 November 2016, Supporting Document: https://www.icann.org/en/system/files/files/proposal-georgian-lgr-24nov16-en.pdf

[Proposal-Khmer] Khmer Generation Panel, "Proposal for Khmer Script Root Zone Label Generation Rules (LGR)", 15 August 2016, https://www.icann.org/en/system/files/files/proposal-khmer-lgr-15aug16-en.pdf

[Proposal-Lao] Lao Generation Panel, "Proposal for a Lao Script Root Zone LGR", January 31, 2017, Supporting Document: https://www.icann.org/en/system/files/files/proposal-lao-lgr-31jan17-en.pdf

[Proposal-Thai] Thai Generation Panel, "Proposal for the Thai Script Root Zone LGR", 25 May 2017, Supporting Document: https://www.icann.org/en/system/files/files/proposal-thai-lgr-25may17-en.pdf

[RFC5890] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework", RFC 5890, August 2010.

[RFC5891] Klensin, J., "Internationalized Domain Names in Applications (IDNA): Protocol", RFC 5891, August 2010.

[RFC5892] Faltstrom, P., Ed., "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)", RFC 5892, August 2010.

[RFC5893] Alvestrand, H., Ed., and C. Karp, "Right-to-Left Scripts for Internationalized Domain Names for Applications (IDNA)", RFC 5893, August 2010.

[RFC6912] Sullivan, A., *et al.,* "Principles for Unicode Code Point Inclusion in Labels in the DNS", RFC 6912, April 2013. = IABCP

[RFC7940] Davies, K. and A. Freytag, "Representing Label Generation Rulesets using XML", RFC 7940, August 2016, https://tools.ietf.org/html/rfc7940

[TF-AIDN] Blog, "Task Force for Arabic Script IDNs"; see Appendix H of [Proposal-Arabic] for members that contributed to the development of the Arabic Script LGR Proposal.

[Unicode63] The Unicode Consortium. The Unicode Standard, Version 6.3.0, defined by: "The Unicode Standard, Version 6.3.0", (Mountain View, CA: The Unicode Consortium, 2013. ISBN 978-1-936213-08-5). http://www.unicode.org/versions/Unicode6.3.0/.

[UCD] UAX #44: *Unicode Character Database.* An integral part of The Unicode Standard. Most recent version available from http://www.unicode.org/reports/tr44/. Version 6.3 available as http://www.unicode.org/reports/tr44/tr44-12.html.