Reference Label Generation Rulesets (LGRs) for the Second Level

Publication Date:	5 November 2020
Prepared By:	Pitinan Kooarmornpatana

Public Comment Proceeding	
Open Date:	24 August 2020
Close Date:	15 October 2020
Staff Report Due Date:	5 November 2020

Important Information Links
Announcement
Public Comment Proceeding
View Comments Submitted

Staff Contact:	Pitinan Kooarmornpatana	**Email:**	pitinan.koo@icann.org

Section I: General Overview and Next Steps

To improve the transparency and consistency of the Internationalized Domain Name (IDN) table review process to facilitate the registry operations of new gTLDs, ICANN has developed additional reference IDN Tables in machine-readable format, called reference Label Generation Rulesets (LGRs) for the second level. The reference IDN tables are based on the Guidelines for Developing Reference Label Generation Rules (LGRs), which were finalized after community review. These reference LGRs will be used in reviewing IDN tables submitted by the gTLD registries, e.g. through the Registry Service Evaluation Policy (RSEP) process.

Section II: Contributors

At the time this report was prepared, a total of twelve (12) community submissions had been posted to the forum. The contributors, both individuals and organizations/groups, are listed below in chronological order by posting date with initials noted. To the extent that quotations are used in the foregoing narrative (Section III), such citations will reference the contributor's initials.

Organizations and Groups:

Name	Submitted by	Initials
Chinese Generation Panel and Chinese Domain Name Consortium	Wei WANG, CGP Co-Chair	CGP-CDNC
Thai Generation Panel	Wanawit Ahkuputra, Chair	THGP
Khmer Generation Panel	Rapid Sun, Secretary	KHGP
Neo-Brahmi Generation Panel	Udaya Narayana Singh, Co-Chair	NBGP
Lao National Internet Center	Anisone Kingsada, Director of ccTLD .LA Division	LANIC
Ethiopic Generation Panel	Dessalegn Yehuala, Chair	EGP
Task Force for Arabic Script IDNs	Nabil Beneamar, Member	TF-AIDN

Name	Submitted by	Initials
gTLD Registries Stakeholder Group	Samantha Demetriou, Vice Chair	RySG
At-Large Advisory Committee	ICANN Policy Staff in Support of the At-Large Community	ALAC

Individuals:

Name	Affiliation (if provided)	Initials
Arthit Suriyawongkul		AS
Akshat Joshi	ThinkTrans LLP	AJ
Bill Jouris		BJ

Section III: Summary of Comments

General Disclaimer: This section intends to summarize broadly and comprehensively the comments submitted to this Public Comment proceeding but does not address every specific position stated by each contributor. The preparer recommends that readers interested in specific aspects of any of the summarized comments, or the full context of others, refer directly to the specific contributions at the link referenced above (View Comments Submitted).

CGP-CDNC: CGP-CDNC agrees with the published Chinese reference LGR for the second level and informs that CGP-CDNC will launch an internal process to discuss whether it is needed to merge the four existing sets of rules for Chinese domain labels: Root Zone LGR, Second Level LGR (variant sets based on .asia ZH set), CDNC rules, and .asia rules.

THGP: THGP confirms that the Reference LGR for the second level correctly represents Thai Generation Panel's proposal.

KHGP: KHGP supports the Khmer script LGR for the second level.

NBGP: NBGP informs that there are no further comments to the reference LGRs. It suggests that the confusable cases should be captured in the descriptive part of the LGR.

LANIC: LANIC agrees with the reference LGRs which includes Lao digits.

EGP: EGP has reviewed the Ethiopic script reference LGR for the second level and does not have any additional comments.

TF-AIDNs: TF-AIDNs thanks ICANN org for publishing the Arabic language reference LGR for second level and confirms that the reference LGR agrees with TF-AIDNs' recommendations.

RySG: RySG submits the following comments:

RySG1. The disproportionate burden on IDN implementation could stifle adoption by registries which ultimately limits distribution by registrars and other registration channels.

The additional rules beyond the normative IDNA2008 ranges from 3 to 14 rules combined in 17 reference LGRs. Also, rules vary greatly with 14 rules for Malayalam vs. none for the Chinese LGR. RySG asks what severe security and stability issues these rules are solving for, why now, and why the number of rules are greatly different across scripts.

RySG2: RySG appreciates the effort to improve predictability and consistency in IDN table review process. Creating and enforcing policy within contractual agreement is outside of the bottom up consensus building policy process. The contractual requirements regarding the IDN implementation are clearly stipulated in the registry agreement and in the ICANN IDN Guidelines. Conformance to ICANN reference LGRs for the second level can be encouraged but should not be required, because the work product was not the result of consensus policy. If ICANN org or at-large ICANN community believes these issues need to be discussed. RySG looks forward to working with ICANN org and other stakeholders for the operational framework in accordance with contractual and consensus policies.

RySG3: ICANN policies for registries should not be expected to solve for hardware and/or software limitations which registry operators cannot control. For example; the international reachability, which is not a problem specific for Arabic language, the same issue can be found in other languages (e.g. French, German or Spanish users using an English-only keyboard). This could be potential feature a registry operator might want to offer to address its market's needs, but it should not be expected or required to be implemented.

RySG4: The domain name system should not need to conform to the 'nature' of languages or scripts. Domain name can be reasonable and memorable mnemonics without imposing complex rules to replicate the use of language or script i.e. syllable structure of scripts. This could be potential feature a registry operator might want to offer as a value-added service, but it should not be expected or required to be implemented.

RySG5: RFC5891 does not prohibit ASCII only labels. RySG suggests that the ASCII-only label should not be determined as invalid.

RySG concludes that the IDN policies should (1) be simple and scalable to foster adoption and growth, (2) be focused on addressing the most egregious forms of potential security and stability risks (i.e. in-label script mixing, whole-script confusable labels), and (3) be developed in accordance with contractual and consensus policies to ensure greater acceptance by the ICANN community.

ALAC: ALAC ratifies and submits the ALAC statement on Reference Label Generation Rulesets (LGRs) for the second level in four areas:

ALAC1: Where the documentation uses the "existing registry practice", the links should be provided.

ALAC2: ALAC notes that as the work is based on the root-zone LGR, there are code points outside of the Maximal Starting Repertoire (MSR) which have not been evaluated. These code points should be included in the evaluation as well.

ALAC3: The criteria used by the Generation Panels based on the LGR procedure are extremely narrow as the additional cases would be evaluated by a Similarity Review Panel. No such panel

is envisioned and not possibly practical for the second level. Therefore, ALAC believes that the definition of variants for SLDs needs to be substantially expanded.

ALAC4: ALAC requests revising the text in the document to read: "gTLD registry operators *will* incorporate these reference LGRs when they design their IDN variant tables". ALAC believes that, as something that is critical for security and the avoidance of DNS Abuse, blocking variant names should not be optional.

AS: Phinthu (U+0E3A) is a small diacritic below a consonant which could be a challenge to see on a small browser's address bar or when it is with the consonant with 'tail'. Therefore, two labels, one with and one without Phinthu, could be confusingly similar. The Maiyamok (U+0E46) is the repetition mark. A repeating label can be fully spelt out which will remove the need of the repetition mark and can create ambiguity.

AJ: AJ thanks ICANN org for the reference LGRs and suggests that the Root Zone LGR, which is the base of the work, does not indicate the "similar looking" code points. The String Similarity Assessment Panel exists for the root level string assessment but does not exist for the second and higher levels. Therefore, AJ strongly recommends that ICANN commission a separate study to evaluate the similar looking cases. AJ notes that the Uniform Domain-Name Dispute-Resolution Policy (UDRP) is a mechanism after the dispute happens which is not a desired outcome from the end-user's point of view. AJ provides some representative examples of similar-looking cases for Devanagari script.

BJ: BJ notes that there are several cases where ASCII numerals and scripts numerals are variants. However, BJ notes that are also non-numeral code points which are visually similar to ASCII numerals. For examples: Numeral 0 (U+0030) and Ethiopic Syllable Pharyngeal A ዐ (U+12D0), Numeral 3 (U+0033) and Latin Small Letter Ezh ʒ (U+0292).

Section IV: Analysis of Comments

General Disclaimer: This section intends to provide an analysis and evaluation of the comments submitted along with explanations regarding the basis for any recommendations provided within the analysis.

ICANN org thanks all contributors for their valuable input and feedback.

CGP-CDNC, EGP, KHGP, LANIC, NBGP, THGP, and TF-AIDNs support the reference LGRs.

RySG1. The whole-label and context rules for these LGRs reflect the complex nature of the scripts covered, as has been communicated by the relevant script community, in addition to the restrictions that are part of the IDNA2008. These additional considerations are expected by IDNA2008 and elaborated in RFC6912. Based on RFC5980, registries at all levels of the DNS, not just the top level, are expected to establish policies about label registrations. RFC5984 recommends that registries should develop and apply additional restrictions as needed to reduce confusion and other problems. The rules included provide additional constraints for the script to address the principles of Least Astonishment and Contextual Safety in RFC6912.

The reference LGRs are developed based on the guidelines and the solutions provided by the script communities which have become available through the work on root-zone LGR project (RZ-LGR). In designing the minimally required rules, each script-community has already attempted to balance the complexity introduced by incorporating these rules with the script requirements, following the Simplicity principle (in RFC6912) and provided a conservative solution.

The variation in the set of rules across scripts depends on how different scripts are structured, and where some writing systems are inherently context sensitive (and thus may require rules), others are largely context-free (and may not require rules). For example, Malayalam script is an Abugida writing system with internal structure whereas Chinese is ideographic and context-free on the encoding level. They can therefore expect to show significant differences in the number of rules needed to formulate a label.

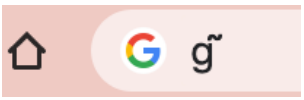
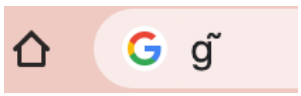
Following the Conservatism Principle in RFC6912, the solutions in the reference LGRs provide a prudent starting point which can be updated based on the additional community feedback. RySG is encouraged to identify any specific rules which it considers need to be relaxed or removed without an impact on security and to communicate these along with the rationale to ICANN org at any time. ICANN org will take any such feedback from RySG and consult with the relevant script community. Based on the conclusion of such discussions, ICANN org would update the reference LGR as needed in the future.

RySG2. Reference LGRs encode information based on input from the relevant script-using community. ICANN org intends to make these available to registries to help them design their own IDN tables and would also use these for IDN table review process. Though reference LGRs provide an overall guidance, adherence to any part of reference LGRs will only be evaluated in the context of security or stability issues. Beyond security and stability issues, adherence to the reference LGRs is not required.

RySG3. Languages using the script like Arabic may have a considerable number of variant labels. As per SSAC's SAC060 report, manageability due to too many allocatable variant labels is also a concern to be addressed. Therefore, the number of allocatable variant labels would need to be constrained. While constraining the number of allocatable variant labels, one would need to ensure that the solutions remain usable across the communities using the language and script.

Some rules, e.g. the international reachability rules in Arabic language LGR, have been proposed by the community to help in limiting the number of allocatable variant labels while balancing it with the usability. However, in designing their own IDN tables, the registries may consider alternative mechanisms to address such issues e.g. by blocking all variant labels. The description in the Arabic language LGR will be enhanced to reflect this detail.

RySG4. The reference LGRs aim to allow labels that represent useful mnemonics which are secure and stable. For example, the Combining Tilde (U+0303) could result in an unpredictable behavior if duplicated, as shown in the following table.

Case	Code points	Screen shot	Browser and OS information
1	U+0067 U+0303 U+0303		Safari Version 14.0 (15610.1.28.1.9, 15610) on macOS Catalina Version 10.15.7
2	U+0067 U+0303 U+0303		Google Chrome Version 86.0.4240.80 (Official Build) (x86_64) on macOS Catalina Version 10.15.7
3	U+0067 U+0303		Google Chrome Version 86.0.4240.80 (Official Build) (x86_64) on macOS Catalina Version 10.15.7

Case 1 and case 2 are the same code point sequence but could be rendered differently based on browser and operating system (OS). In addition, the labels in case 2 and case 3 are not distinguishable (because the duplicate tilde is overlayed in case 2) even though they represent different sequences. This and similar issues with combining marks can be prevented by excluding the singleton U+0303 from the repertoire and only including the required sequence g̃ (U+0067 U+0303). However, in many complex scripts there may be large number of such combinations (possibly in thousands) and so it is more reasonable to handle such cases by rules instead of listing all sequences.

In addition, for complex scripts, ignoring the syllable structure is not a feasible option. These scripts are written and read by native users as syllables, with font engines supporting syllable layout features. For linguistic and technical reasons Unicode does not support the syllables directly for all scripts but instead encodes the underlying characters. Care is being taken in designing these rules not to try to enforce spelling, but only the underlying structure of the writing system. Invalid syllable structure is prevented as it can present unanticipated issues, e.g. as shared in the table above.

RySG5. The rule which limits ASCII-only labels was originally motivated by RFC5890 which states that “A ‘U-label’ is an IDNA-valid string of Unicode characters, in Normalization Form C (NFC) and including at least one non-ASCII character, expressed in a standard Unicode Encoding Form (such as UTF-8)”. Based on RySG’s input, the rule which limits ASCII-only labels will be removed from reference LGRs and the statement noting this change will be added in the description section.

ALAC1. The work has reviewed multiple IDN tables published in the IANA repository. Where reference LGRs rely on existing IDN tables, these have been cited (in the XML/HTML) and where they rely on existing RZ-LGR tables, these tables cite any existing practice that was relied on in their design.

ALAC2. Code points currently not included in script reference LGRs tend to generally be ones that are in limited use, which in many cases may mean that they are not found on widely standardized keyboards or are unfamiliar to large subsets of the user base. The current solution offers a conservative secure and stable solution on code points which have already been reviewed by the script community and considered sufficient to represent the common usage of the script.

Reference LGRs can be updated over time and additional code points can be added in the reference LGRs in the future, based on further analysis and input from the script community.

This constraint, however, still does not limit a registry operator which may still add additional code points in its IDN tables even if these are not included in the reference LGR, as long as such inclusion does not cause a security or stability issue.

ALAC3 and AJ. Variant code points are limited to those which represent the “same” or indistinguishable code point as per their definition. Variant relations are generally transitive and symmetric. Variant code points have been determined by the relevant script communities and included in the reference LGRs. These are different from cases which are confusable though distinguishable for a user (which may not be a transitive or symmetric relation). Different mechanisms exist to deal with the latter cases, e.g. Uniform Domain-Name Dispute-Resolution Policy (UDRP), Uniform Rapid Suspension (URS). In case ALAC would like to consider additional variant code point candidates, it is requested that ALAC identifies the specific cases and share with ICANN org, which can then be discussed by the relevant script community.

ALAC4. The application of the reference LGRs is based on existing policies and processes. This already includes review of IDN tables designed by the registry operators for security and stability, while allowing flexibility for the registry operators to suit their business needs. The review addresses the concerns raised by ALAC regarding security.

AS. The comments for Phinthu (U+0E3A) and Maiyamok (U+0E46) are well noted. The explanation is already included in the Thai RZ-LGR supporting document (section 5.4 and section 5.6) which is referred to from the reference LGR.

BJ. The variant sets have been developed based on the confirmation by the relevant script community. In case that additional variant code point candidates should be considered, it is requested that BJ identifies the specific sets and share with ICANN org, which then can be discussed by the relevant script community. Further, for similarity cases, please see response for ALAC3 and AJ.