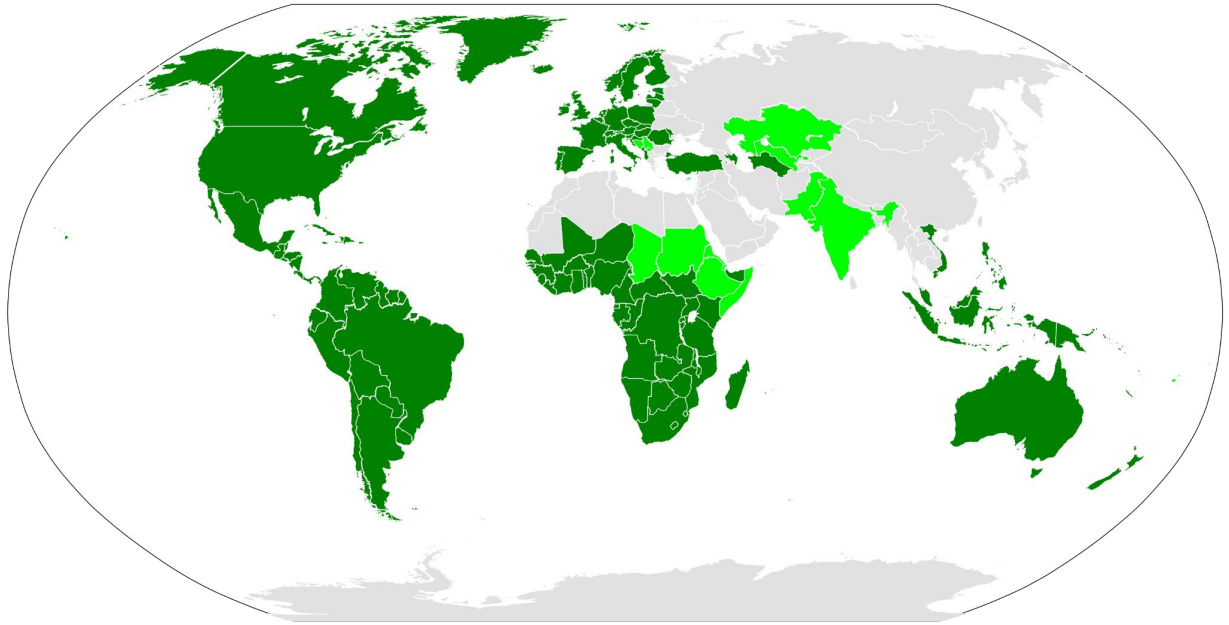
Diacritics Issue in Latin Script

Background Briefing

ICANN78 GNSO Council Open Session | Wednesday, 25 October 2023

Latin Script Basics

- Latin script is a major writing system in the world and the **most widely used** in terms of the number of languages and speakers
  - About 70% of the world's literate population use the Latin script
- 1,189 languages use the Latin script
  - European language examples: Danish, Dutch, English, French, German, Italian, Spanish, etc.
  - Non-European language examples: Chamorro, Filipino, Guarani, Kiribati, Niuean, Turkish, Swahili, Vietnamese, etc.
- 212 languages are considered in the Latin Script Proposal integrated in the Root-Zone Label Generation Rules (RZ-LGR)
- Latin script is related to Armenian, Cyrillic, and Greek scripts (all derived from Greek)



- *Dark green marks countries where the Latin script is the sole main script*
- *Light green marks countries where Latin co-exists with other scripts*
- *Grey marks areas, in which the Latin script is not used or used unofficially for a second language*
- *Source: https://en.wikipedia.org/wiki/Latin_script*

Diacritics Basics

- Diacritics are **modifiers** surrounding basic letter shapes, generally recognized as **distinct graphic elements** to form new letters
- The main use of diacritics in the Latin script is to **change the sound-values** of the letters to which they are added
  - Noval diacritics are used to express less common **distinctive linguistic features**, such as tone
  - Stacking diacritics are developed for linguistically distinctive features **absent from European languages**
- Some non-Latin scripts also use diacritics, such as Arabic, Greek, Hebrew, and Korean

Latin Diacritics Examples

Danish	æ	å	ø												
French	à	â	æ	ç	é	è	ê	ë	î	ï	ô	œ	ù	û	ü	ÿ
German	ä	ö	ü	ß											
Latvian	ā	č	ē	ģ	ī	ķ	ļ	ņ	š	ū	ž				
Spanish	á	é	í	ñ	ó	ú	ü								
Swedish	å	ä	ö												
Turkish	ç	ğ	ı	ö	ş	ü									
Vietnamese	ă	â	đ	ê	ô	ơ	ư	à	á	ả	ã	ạ			

Diacritics Omission

- It's a common practice to omit diacritics by making the skeleton form of a word understandable or workable
- In the DNS context, omitting diacritics turns a label into an ASCII base when IDNs were not supported; users have adapted accordingly, much the same way they all adapted to the absence of spaces in domain names

French	.déjà	.deja
Spanish	.mañana	.manana
Portuguese	.violão	.violao

  - It can still become awkward as when reading ".thisdomainwayistoolong"
- In some cases, like German, the language has adapted to tolerate the alternate spelling in basic Latin to serve foreign audiences
  - Example: **Köln < > Koeln**
- **However, diacritic omission does not necessarily mean the ASCII base is the Equivalence of the original version in its respective language**
  - Native speakers would not naturally or correctly write the word without diacritics
  - The ASCII base is not officially recognized as correct in a given language, but rather a "shortcut" or "workaround"
  - As such, "deja", "manana", and "violao" would not be perceived as correct in their respective languages

Variant Basics

- **Variant means the correct alternative alphabet / character / label that differ in some respect and form but mean exactly the same thing according to the rules of a given language**
- Variants exist in many scripts to serve language communities globally, impacting billions of users; a single script (e.g., Latin) can be used in multiple languages and may be subject to variations due to how the languages work
- DNS makes distinctions between variants with different code points, but script community recognizes them as being equivalent; variants may exacerbate confusion risks among labels that may or may not be visually similar
- **RZ-LGR (latest version 5)** offers way to have consistent definitions, at the TLD level, for variants in 25 scripts
- Both SubPro PDP and EPDP-IDNs have affirmed the **authoritative status of RZ-LGR** for defining variant gTLDs and determining whether they are allocatable or blocked

Arabic Script	**Chinese Script**
لاسعودية (Arabic Language)	长城 (Simplified Chinese)
لاسعودية (Urdu Language)	長城 (Traditional Chinese)
Saudi Arabia (alsauduit)	*Great Wall*

Scripts with Variants

- Arabic
- Armenian
- Bangla (Bengali)
- Chinese (Han)
- Cyrillic
- Devanagari
- Ethiopic
- Georgian
- Greek
- Gujarati
- Gurmukhi
- Hebrew
- Japanese
- Kannada
- Khmer
- Korean
- Lao
- Latin
- Malayalam
- Myanmar
- Oriya
- Sinhala
- Tamil
- Telugu
- Thaana
- Tibetan
- Thai

- Variant - 22 scripts
- Allocatable variant - 7 scripts
- No variant - 4 scripts
- Work in progress - 2 scripts

Example Output Using RZ-LGR

- **Original**: Primary / source label
- **Allocable**: Available for delegation but must be applied for
- **Blocked**: Unavailable for delegation

#	Type	U-label	A-label	Disposition	Code point sequence
1	original	شبكة	xn--ngbc5azd	valid	U+0634 U+0628 U+0643 U+0629
2	varlabel	شبكه	xn--ngbx0cq	allocatable	U+0634 U+0628 U+0643 U+0647
3	varlabel	شبكھ	xn--ngbx0c15a	blocked	U+0634 U+0628 U+0643 U+06BE
4	varlabel	شبكۀ	xn--ngbx0c95a	blocked	U+0634 U+0628 U+0643 U+06C0
5	varlabel	شبكہ	xn--ngbx0cy6a	blocked	U+0634 U+0628 U+0643 U+06C1
6	varlabel	شبكۂ	xn--ngbx0c26a	blocked	U+0634 U+0628 U+0643 U+06C2
7	varlabel	شبكۃ	xn--ngbx0c66a	allocatable	U+0634 U+0628 U+0643 U+06C3
8	varlabel	شبكه	xn--ngbx0c31b	blocked	U+0634 U+0628 U+0643 U+06D5
9	varlabel	شبكة	xn--ngbc5az1b	allocatable	U+0634 U+0628 U+06A9 U+0629
10	varlabel	شبكه	xn--ngbx2d5u	allocatable	U+0634 U+0628 U+06A9 U+0647
11	varlabel	شبكھ	xn--ngbx66ayc	blocked	U+0634 U+0628 U+06A9 U+06BE
12	varlabel	شبكۀ	xn--ngbx66a6c	blocked	U+0634 U+0628 U+06A9 U+06C0
13	varlabel	شبكہ	xn--ngbx66agd	blocked	U+0634 U+0628 U+06A9 U+06C1
14	varlabel	شبكۂ	xn--ngbx66akd	blocked	U+0634 U+0628 U+06A9 U+06C2
15	varlabel	شبكۃ	xn--ngbx66aod	allocatable	U+0634 U+0628 U+06A9 U+06C3
16	varlabel	شبكه	xn--ngbx66a0f	blocked	U+0634 U+0628 U+06A9 U+06D5
17	varlabel	شبكة	xn--ngbc5a31b	allocatable	U+0634 U+0628 U+06AA U+0629
18	varlabel	شبكه	xn--ngbx2d9u	allocatable	U+0634 U+0628 U+06AA U+0647
19	varlabel	شبكھ	xn--ngbx96asc	blocked	U+0634 U+0628 U+06AA U+06BE
20	varlabel	شبكۀ	xn--ngbx96a0c	blocked	U+0634 U+0628 U+06AA U+06C0
21	varlabel	شبكہ	xn--ngbx96a4c	blocked	U+0634 U+0628 U+06AA U+06C1
22	varlabel	شبكۂ	xn--ngbx96a8c	blocked	U+0634 U+0628 U+06AA U+06C2
23	varlabel	شبكۃ	xn--ngbx96ahd	allocatable	U+0634 U+0628 U+06AA U+06C3
24	varlabel	شبكه	xn--ngbx96arf	blocked	U+0634 U+0628 U+06AA U+06D5

Latin Variants in RZ-LGR

- Variants exist in the Latin script of the Root-Zone Label Generation Rules
- Latin Generation Panel (GP) defined variants based on:
  - Exactly identical shapes
  - Letter shapes that will be misidentified (unless the reader could tell a different language context intended)
  - Consideration of Armenian, Cyrillic, and Greek scripts due to overlap in letter shapes
- **Variants in the Latin script are generally blocked** to mitigate potential user confusion
- **Diacritic letters are generally NOT defined as variants of their ASCII base because they are deemed distinguishable**
- Due to integration of the Armenian, Cyrillic, and Greek scripts, some variant pairings do exist between certain ASCII base letters and diacritic letters. However, they are **blocked** (examples)

0061	a	00E1	á	↔	**blocked**
0069	i	00EF	ï	↔	**blocked**

006E	n	0144	ń	↔	**blocked**
006F	o	00F3	ó	↔	**blocked**

- **RZ-LGR has placed strict limitations on the instances of allocatable variant paring with only two exceptions**

German	**“ß”** used in Standard German	Allocatable *(in one direction)*	**“ss”** used in Basic Latin
Turkish	**“ı”** dotless i used in Turkish		**“i”** used in Basic Latin

Problem Statement

An IDN gTLD with diacritics may be unlikely to co-exist with its base ASCII gTLD

What is the reason?

- A correctly spelt IDN with diacritics is likely either a non-variant to its shortcut ASCII base, or a blocked variant
- In the “non-variant” scenario, even the Latin GP deems the IDN and its ASCII base “distinguishable”, they may still be determined confusingly similar during String Similarity Review, as the diacritics only add a small change to the ASCII base letter shape and may lead to user confusion
- Since many existing gTLDs omitted the diacritics to adapt to the DNS, their correctly spelt IDN versions, if applied for, will likely face such a challenge during String Similarity Review and may be ineligible to proceed
- If an IDN gTLD already exists, its ASCII base version without the diacritics, if applied for, will likely face the same challenge during String Similarity Review

What is the implication?

- An existing registry or a new applicant may face the dilemma of choosing between a shortcut **ASCII string** for globalization or accessibility purposes and a correctly spelt **IDN string** for localization or identity reasons

Impact on Existing gTLDs

	Existing gTLD from 2012 Round	Potential Applied-for String
French	.hermes	“.hermès”
German	.koeln	“.köln” “.koln”
French	.lancome	“.lancôme”
French	.quebec	“.québec”
German	.vermögensberater	“.vermogensberater” “.vermoegensberater”
German	.vermögensberatung	“.vermogensberatung” “.vermoegensberatung”
German	.zuerich	“.zürich” “.zurich”

Note: these may be the only existing gTLDs in the Latin script that may encounter the diacritics issue

Considerations for Issue Scoping

GNSO policy solution is required to develop an exception to address diacritics issue for existing / future gTLDs

1. **What is the scope of the problem?**
    a. Only the languages with diacritics in the Latin script? (basic Latin consists of ASCII letters)
    b. Applied-for IDN strings of existing ASCII gTLDs that work as a "shortcut"? Applied-for ASCII strings that act as a "shortcut" of existing IDN gTLDs?
    c. Brand new applications?
    d. Limit to Geographic Names TLDs? Or expand to .brand TLDs, community TLDs, and other types? (e.g., ".HäagenDazs", ".Røde", ".MötleyCrüe", ".Motörhead", etc.)
2. **What should be excluded from consideration?**
    a. Singular / plural, in English and other languages?
    b. Alternative spellings? (e.g., program / programme, analyze / analyse, Mexico / Mejico)
    c. Blocked variants as calculated by RZ-LGR?
3. **What is the criteria for establishing "equivalence"?**
    a. Is it based on dictionary definition for a given language?
    b. If the dictionary cannot verify "equivalence", what other proof would suffice? (e.g., a registered mark in a given language)?

Considerations for Issue Scoping (Cont.)

4. **What are the considerations for a potential solution?**
   a. Should all of the EPDP-IDNs Phase 1 recommendations apply from the application, contractual, and operational standpoint?
      i. Notion of “primary” and “equivalent” strings? Which string is the “primary”, ASCII or IDN?
      ii. How to apply the “same entity” principle?
   b. Should all of the EPDP-IDNs Phase 2 recommendations apply to second-level domain name registrations under those strings?
   c. To what extent would the ccTLD solution be borrowed / referenced?

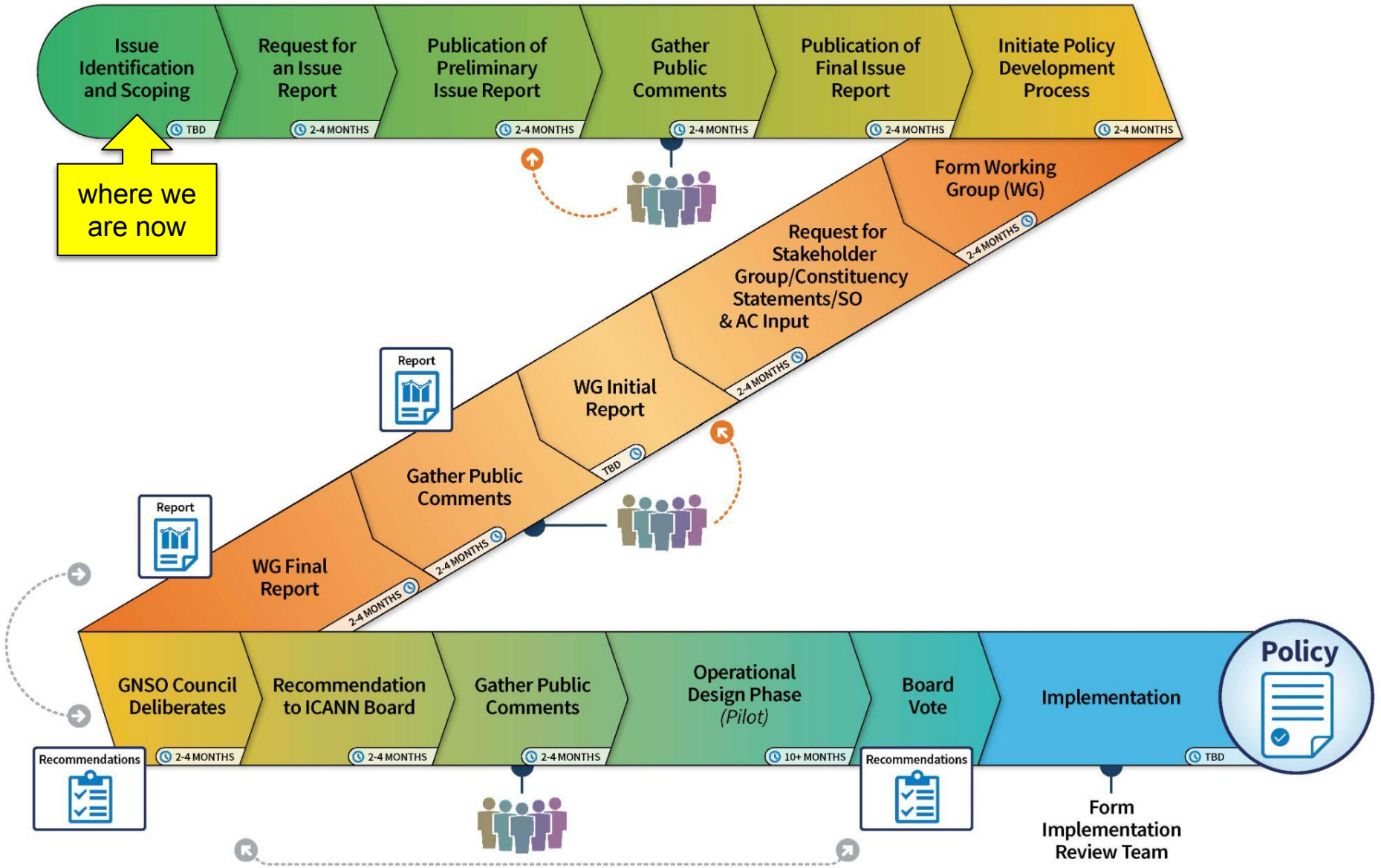
Conservatism Principle:

- This principle advocates for the adoption of a more cautious approach as a way to limit any potential security and stability risks associated with the gTLD string delegation in the absence of data or information in support of a more liberal approach. It is consistent with RFC 6912 which says, “doubts should always be resolved in favor of rejecting”.
- An exception should be minimal in scope

Resources for Further Reading

- RZ-LGR-Version 5 Overview, Section 2.3.19, Latin LGR Proposal
  https://www.icann.org/sites/default/files/lgr/rz-lgr-5-overview-26may22-en.pdf
- Latin LGR HTML version: https://www.icann.org/sites/default/files/lgr/rz-lgr-5-latin-script-26may22-en.html
- Latin LGR Supporting Documents: https://www.icann.org/en/system/files/files/proposal-latin-lgr-23sep21-en.pdf
- Latin LGR Appendices: https://www.icann.org/en/system/files/files/proposal-latin-lgr-appdendices-23sep21-en.zip
- Discussion Paper About “.québec” Challenges:
  https://mm.icann.org/pipermail/council/attachments/20230817/035c9620/DiscussionPaperAbout.qubecChallenges-0001.pdf

Next Steps


Issue Identification and Scoping
TBD
Request for an Issue Report
2-4 MONTHS
Publication of Preliminary Issue Report
2-4 MONTHS
Gather Public Comments
2-4 MONTHS
Publication of Final Issue Report
2-4 MONTHS
Initiate Policy Development Process
2-4 MONTHS
where we are now
Form Working Group (WG)
2-4 MONTHS
Request for Stakeholder Group/Constituency Statements/SO & AC Input
2-4 MONTHS
Report
WG Initial Report
TBD
Gather Public Comments
2-4 MONTHS
Report
WG Final Report
2-4 MONTHS
GNSO Council Deliberates
2-4 MONTHS
Recommendations
Recommendation to ICANN Board
2-4 MONTHS
Gather Public Comments
2-4 MONTHS
Operational Design Phase (Pilot)
10+ MONTHS
Board Vote
Recommendations
Implementation
TBD
Policy
Form Implementation Review Team