IDN TLD Program at ICANN

TLDCON, Baku, Azerbaijan

Thursday, 11 September 2014

Top Level Domain (TLD) Labels

www.icann.org

- Traditionally rules for a label
  - ASCII Letters [a-z], Digits [0-9] and Hyphen (LDH)
  - Maximum length = 63

- Traditionally rules for TLD label
  - Letter Principle: ASCII [a-z] only
  - Maximum length = 63 (in practice much shorter)
    - Two characters for ccTLDs - .az, .bg, .mn, .ru, …
    - Longer labels for gTLDs – .com, .org, .info, .museum, …

ICANN

Internationalized Domain Name (IDN) TLD labels

- Rules for IDN Labels
  - Valid U-Label
    - Only containing PVALID and CONTEXT O/J code points - IDNA 2008
    - Constraints on mixing scripts
  - Valid A-Label
    - "xn--" + Punycode of U-label - RFC 3492
    - Maximum A-label length = 63 = "xn--" + 59
    - Maximum Punycode length = 59

- Rules for IDN TLD Labels
  - "Letter" principle for U-Label
  - Valid A-Label
    - Maximum A-label length = "xn--" + 59

ICANN

Rules for IDN TLDs

1. Which characters can form a label?

2. Which characters (if any) are confusable or variants?

3. What are additional constraints on labels?

Which Characters are Needed for TLDs?

- Must be in common used to write a language
  - Must be in widespread modern use, e.g. not historical characters
    - Cyrillic
    - Cyrillic Supplement
    - Cyrillic Extended-A
    - Cyrillic Extended-B
  - Limit based on languages in RFC 5992 and others (Mongolian)

- Preferably adhere to *Principles for Unicode Code Point Inclusion in Labels in the DNS* - RFC 6912

Code Points

	040	041	042	043	044	045	046	047	048	049	04A	04B	04C	04D	04E	04F	050	051	052
0	Ѐ 0400	А 0410	Р 0420	а 0430	р 0440	ѐ 0450	Ѡ 0460	Ѱ 0470	Ҁ 0480	Ґ 0490	Ҡ 04A0	Ұ 04B0	Ӏ 04C0	Ӑ 04D0	Ӡ 04E0	Ӱ 04F0	Ԁ 0500	Ԑ 0510	Ԡ 0520
1	Ё 0401	Б 0411	С 0421	б 0431	с 0441	ё 0451	ѡ 0461	ѱ 0471	ҁ 0481	ґ 0491	ҡ 04A1	ұ 04B1	Ӂ 04C1	ӑ 04D1	ӡ 04E1	ӱ 04F1	ԁ 0501	ԑ 0511	ԡ 0521
2	Ђ 0402	В 0412	Т 0422	в 0432	т 0442	ђ 0452	Ѣ 0462	Ѳ 0472	҂ 0482	Ғ 0492	Ң 04A2	Ҳ 04B2	ӂ 04C2	Ӓ 04D2	Ӣ 04E2	Ӳ 04F2	Ԃ 0502	Ԓ 0512	Ԣ 0522
3	Ѓ 0403	Г 0413	У 0423	г 0433	у 0443	ѓ 0453	ѣ 0463	ѳ 0473	҃ 0483	ғ 0493	ң 04A3	ҳ 04B3	Ӄ 04C3	ӓ 04D3	ӣ 04E3	ӳ 04F3	ԃ 0503	ԓ 0513	ԣ 0523
4	Є 0404	Д 0414	Ф 0424	д 0434	ф 0444	є 0454	Ѥ 0464	Ѵ 0474	҄ 0484	Ҕ 0494	Ҥ 04A4	Ҵ 04B4	ӄ 04C4	Ӕ 04D4	Ӥ 04E4	Ӵ 04F4	Ԅ 0504	Ԕ 0514	Ԥ 0524
5	Ѕ 0405	Е 0415	Х 0425	е 0435	х 0445	ѕ 0455	ѥ 0465	ѵ 0475	҅ 0485	ҕ 0495	ҥ 04A5	ҵ 04B5	Ӆ 04C5	ӕ 04D5	ӥ 04E5	ӵ 04F5	ԅ 0505	ԕ 0515	ԥ 0525
6	І 0406	Ж 0416	Ц 0426	ж 0436	ц 0446	і 0456	Ѧ 0466	Ѷ 0476	҆ 0486	Җ 0496	Ҧ 04A6	Ҷ 04B6	ӆ 04C6	Ӗ 04D6	Ӧ 04E6	Ӷ 04F6	Ԇ 0506	Ԗ 0516	Ҍ 0526
7	Ї 0407	З 0417	Ч 0427	з 0437	ч 0447	ї 0457	ѧ 0467	ѷ 0477	҇ 0487	җ 0497	ҧ 04A7	ҷ 04B7	Ӈ 04C7	ӗ 04D7	ӧ 04E7	ӷ 04F7	ԇ 0507	ԗ 0517	ҍ 0527
8	Ј 0408	И 0418	Ш 0428	и 0438	ш 0448	ј 0458	Ѩ 0468	Ѹ 0478	҈ 0488	Ҙ 0498	Ҩ 04A8	Ҹ 04B8	ӈ 04C8	Ә 04D8	Ө 04E8	Ӹ 04F8	Ԉ 0508	Ԙ 0518	
9	Љ 0409	Й 0419	Щ 0429	й 0439	щ 0449	љ 0459	ѩ 0469	ѹ 0479	҉ 0489	ҙ 0499	ҩ 04A9	ҹ 04B9	Ӊ 04C9	ә 04D9	ө 04E9	ӹ 04F9	ԉ 0509	ԙ 0519	
A	Њ 040A	К 041A	Ъ 042A	к 043A	ъ 044A	њ 045A	Ѫ 046A	Ѻ 047A	Ҋ 048A	Қ 049A	Ҫ 04AA	Һ 04BA	ӊ 04CA	Ӛ 04DA	Ӫ 04EA	Ӻ 04FA	Ԋ 050A	Ԛ 051A	
B	Ћ 040B	Л 041B	Ы 042B	л 043B	ы 044B	ћ 045B	ѫ 046B	ѻ 047B	ҋ 048B	қ 049B	ҫ 04AB	һ 04BB	Ӌ 04CB	ӛ 04DB	ӫ 04EB	ӻ 04FB	ԋ 050B	ԛ 051B	
C	Ќ 040C	М 041C	Ь 042C	м 043C	ь 044C	ќ 045C	Ѭ 046C	Ѽ 047C	Ҍ 048C	Ҝ 049C	Ҭ 04AC	Ҽ 04BC	ӌ 04CC	Ӝ 04DC	Ӭ 04EC	Ӽ 04FC	Ԍ 050C	Ԝ 051C	
D	Ѝ 040D	Н 041D	Э 042D	н 043D	э 044D	ѝ 045D	ѭ 046D	ѽ 047D	ҍ 048D	ҝ 049D	ҭ 04AD	ҽ 04BD	Ӎ 04CD	ӝ 04DD	ӭ 04ED	ӽ 04FD	ԍ 050D	ԝ 051D	
E	Ў 040E	О 041E	Ю 042E	о 043E	ю 044E	ў 045E	Ѯ 046E	Ѿ 047E	Ҏ 048E	Ҟ 049E	Ү 04AE	Ҿ 04BE	ӎ 04CE	Ӟ 04DE	Ӯ 04EE	Ӿ 04FE	Ԏ 050E	Ԟ 051E	
F	Џ 040F	П 041F	Я 042F	п 043F	я 044F	џ 045F	ѯ 046F	ѿ 047F	ҏ 048F	ҟ 049F	ү 04AF	ҿ 04BF	ӏ 04CF	ӟ 04DF	ӯ 04EF	ӿ 04FF	ԏ 050F	ԟ 051F	

Which Code Points are Variants?

- Within Cyrillic script
  - Are characters with and without diacritics confusable?
    - ѝ (U+045D) and и (U+0438)
  - Are there any other confusable characters?
    - г (U+0433) and ґ (U+0491)
- Across script, with Latin and Greek
  - Lower case confusable labels?
    - pear.ae – Cyrillic or Latin?
  - Upper case confusable labels?
    - BET.HM – Cyrllic, Greek or Latin?
  - How to manage such confusable labels?

IDN TLD Program

- To support IDNs and variants in the root zone, the ICANN community, at the direction of the Board[1], undertook several projects to study and make recommendations on their viability, sustainability and delegation

- Being conducted in multiple phases, these projects form the basis of an implementation plan that will be considered by ICANN's Board of Directors

[1] https://www.icann.org/resources/board-material/resolutions-2010-09-25-en

Phases Completed under the IDN Variant Program

PHASE 1 (2011)

Case Studies:
- Arabic
- Chinese
- Cyrillic
- Devanagari
- Greek
- Latin

PHASE 2 (2011-12)

Integrated Issues Report

PHASE 3 (2012-13)

Projects:
- P1 LGR XML Specification
- P2.1 LGR Process for the Root Zone
- P6 User Experience Study for TLD Variants

https://www.icann.org/resources/pages/reports-2013-04-03-en

ICANN

Phase 1: Script Case Studies

- Investigated issues relevant to scripts that need to be resolved to facilitate a good user experience for IDN variant TLDs

  ○ Arabic, Chinese, Cyrillic, Devanagari, Greek and Latin

  ○ 66 experts from 29 countries and territories

  ○ expertise in DNS, IDNs, linguistics, security & scalability, policy, registry/registrar operations

  ○ The reports posted in October 2011

    ▪ Available at: https://www.icann.org/resources/pages/variant-tlds-2012-05-08-en

ICANN
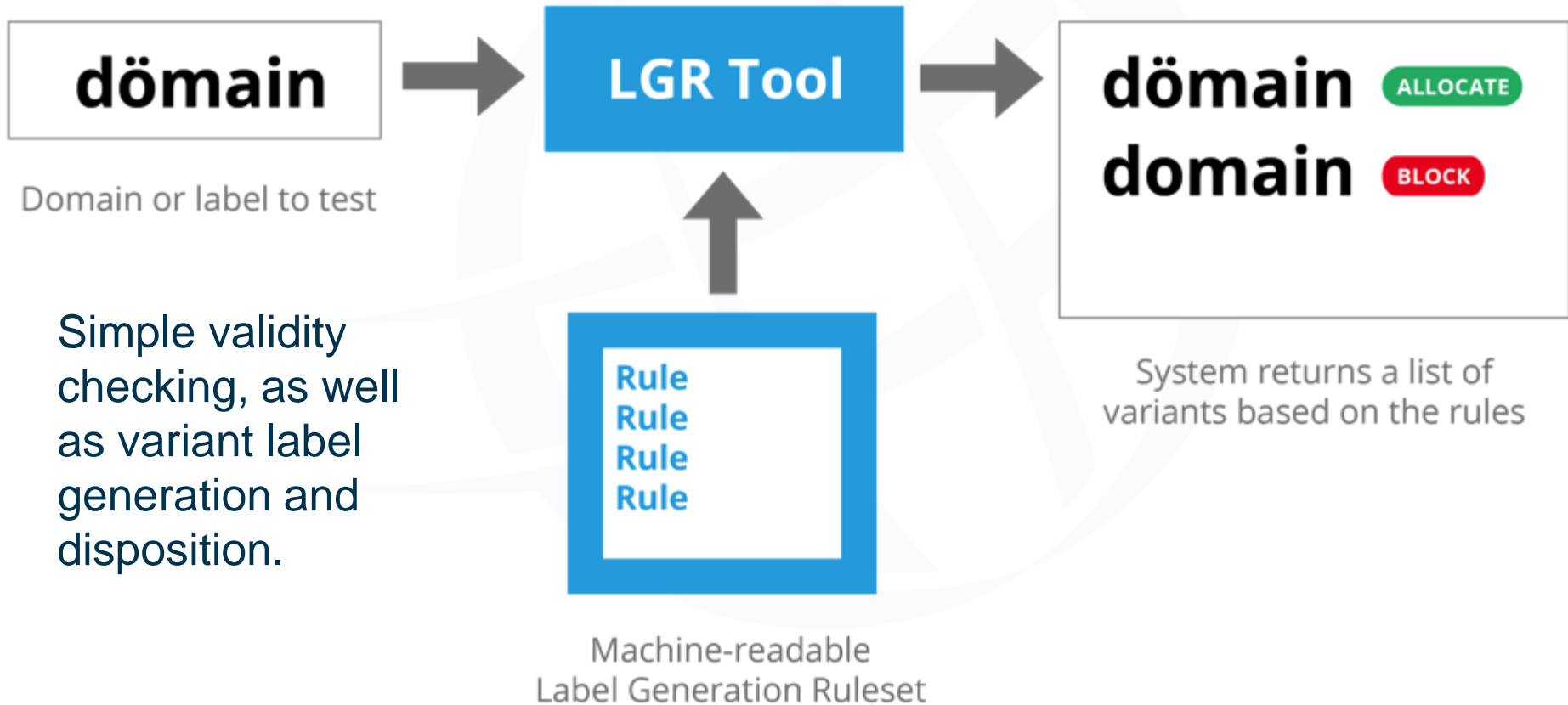
Phase 2: Integrated Issues Report

- The six script case studies used to create a common framework of issues with the possible inclusion of IDN variant TLDs in the DNS root zone
  - ICANN was assisted by a coordination team comprised of representatives from the case study teams
    - Provided valuable language and script support
    - Served as reviewer to the structure and content of this report
  - Developed a working definition of 'variants'
  - Clarified the issues and identified complexities
  - Recommended for additional work on the issues and identification of potential solutions
  - The report posted in December 2011

ICANN

Phase 3: Label Generation Ruleset (LGR) and User Experience Implications

- Based on Phase 2, ICANN, in consultation with the community, identified and prioritized next steps as three projects

  - Label Generation Ruleset Specification and Tool – P1

  - Label Generation Ruleset (LGR) Process for the Root – P2.1

  - Examining the User Experience Implications of Active Variant TLDs – P6

  - The reports and protocol posted in March 2013

Phase 3: Project P1 - Representing LGRs in XML

- ## What Does XML-LGR Enable?

dömain

Domain or label to test

LGR Tool

dömain ALLOCATE

domain BLOCK

System returns a list of variants based on the rules

Simple validity checking, as well as variant label generation and disposition.

Rule
Rule
Rule
Rule

Machine-readable
Label Generation Ruleset

ICANN

Phase 3: Project P2.1 - LGR Procedure

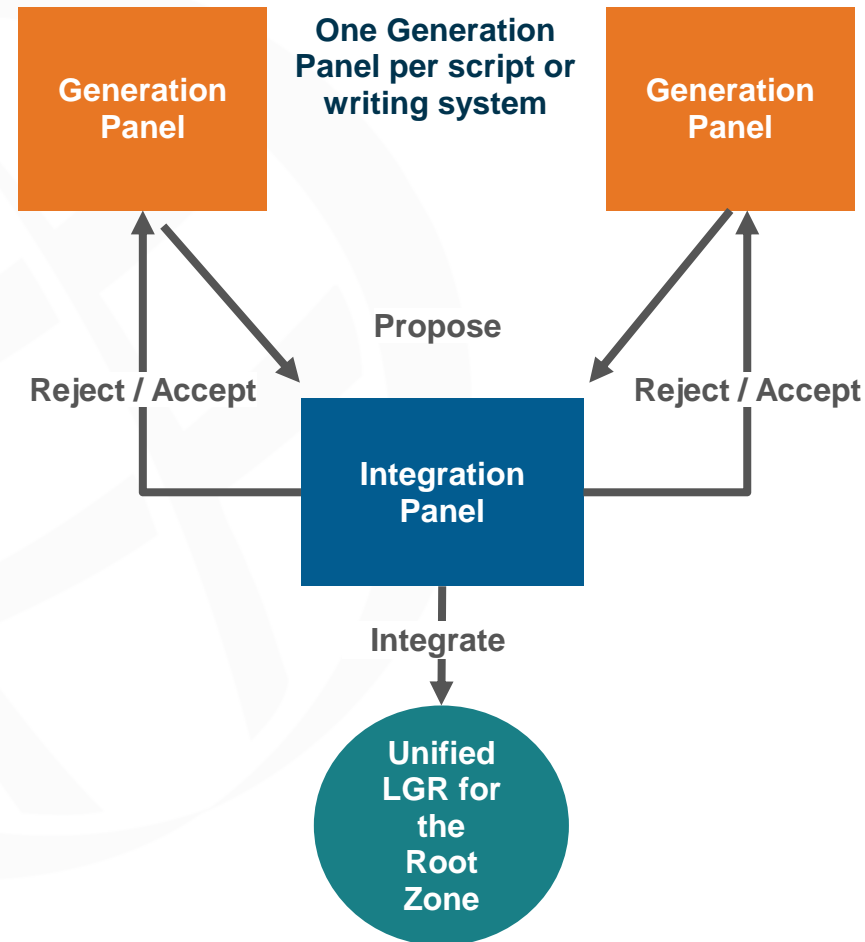**TO BE FORMED BY SCRIPT COMMUNITIES**

Generation Panels

- Generate proposals for script specific LGRs, based on community expertise and requirements

Integration Panel

- Integrates them into common Root Zone LGR while minimizing the risk to Root Zone as shared resource

Label Generation Rules (LGR)

- Which labels are permissible

- Which variant labels exist

- Which variant labels may be allocated

Generation Panel

One Generation Panel per script or writing system

Generation Panel

Propose

Reject / Accept

Reject / Accept

Integration Panel

Integrate

Unified LGR for the Root Zone

ICANN

Phase 3: P6 – User Experience & variant TLDs

- If two or more variants are delegated, what are the user experience issues for the users:

  - End users

  - Power users (system admins, developers, etc.)

  - Registries, registrars, registrants

- Recommendations to provide an acceptable user experience

Current Work: P2.2 Implementation of LGR

- Implementation of the LGR Procedure initiated
  - June 2013 - Call for SMEs for Integration Panel to develop Root Zone LGR
  - July 2013 - Call for Generation Panels to propose Root Zone Label Generation Rules
  - September 2013 - Integration Panel formed
  - February 2014 – Arabic Generation Panel formed
  - March 2014 - MSR-1 released for public comment
  - June 2014 - MSR-1 published
  - September 2014 – Chinese Generation Panel formed
  - Outreach efforts for formulation of additional script Generation Panels

MSR-1 Content in Numbers

- 22 scripts
  - Arabic, Bengali, Cyrillic, Devanagari, Georgian, Greek, Gujarati, Gurmukhi, Han, Hangul, Hebrew, Hiragana, Kannada, Katakana, Lao, Latin, Malayalam, Oriya, Sinhala, Tamil, Telugu and Thai

- 'Common' and 'Inherited' (shared)

- 32,790 code points
  - from 97,973 PVALID/CONTEXT code points defined in Unicode 6.3
  - 11,172 Hangul syllables and 19,850 Han ideographs

ICANN

LGR Procedure Depends on Community Work

- Get involved:

  - Form a generation panel

  - Volunteer to join a generation panel

  - Take part in public review of the MSR, LGR proposals, integrated LGR, etc.

  - Disseminate information to interested communities or individuals

Arabic
Bengali
Chinese
Cyrillic
Devanagari
Georgian
Greek
Gujarati
Gurmukhi
Hebrew
Japanese
Korean
Latin
Sinhala
Tamil
Telugu
Thai

Resources

- Reports and Supporting Material

  o https://www.icann.org/resources/pages/variant-tlds-2012-05-08-en

- Toolkit for 'How to form a Generation Panel'

  o Quick Guide Kit for Generation Panels

- Project mailing lists:

  o LGR@icann.org: Communicate with LGR community members and the Integration Panel on matters related to LGR work

  o IntegrationPanel@icann.org: Contact directly the Integration Panel members on all matters related to LGR work

  o ArabicGP@icann.org, ChineseGP@icann.org, CyrillicGP@icann.org, KoreanGP@icann.org, NeoBrahmiGP@icann.org : script community dedicated mailing lists

  o idntlds@icann.org: Contact ICANN to submit Generation Panel proposals, individual statement of interests, work reports, updates, etc.

  o Discuss issues related to the IDN Variant TLDs Program by subscribing to vip@icann.org here: https://mm.icann.org/mailman/listinfo/vip

ICANN

Thank You

Join us at the IDN Program Sessions

on Wednesday 15 October 2014

at ICANN 51 in Los Angeles