Myanmar Script Generation Panel Meeting

Pitinan Kooarmornpatana
Manager, IDN Programs

Yangon
16-17 June 2018

Agenda

Day 1 – Saturday 16 June, 2018	
0900 – 0910	**Introduction and Welcome to Members** Thin Zar Phyo *Myanmar Generation Panel Chair*
0910 – 0925	**Overview of Enabling Myanmar Language in Computer** Dr. Myint Myint Than *Executive Director & member of Myanmar NLP Working Committee, Myanmar Computer Federation*
0925 – 0930	**Introduction and Welcome to the Myanmar Generation Panel Meeting** Pitinan Kooarmornpatana *IDN Program Manager, ICANN*
0930 – 0950	**Introduction to ICANN and multi-stakeholder model** Joyce Chen *(Senior Manager ICANN Global Stakeholder Engagement Strategy & Development, ICANN)*
0950 – 1010	**Overview of Root Zone Label Generation Rules Program** Pitinan Kooarmornpatana **Photo Session**

Agenda

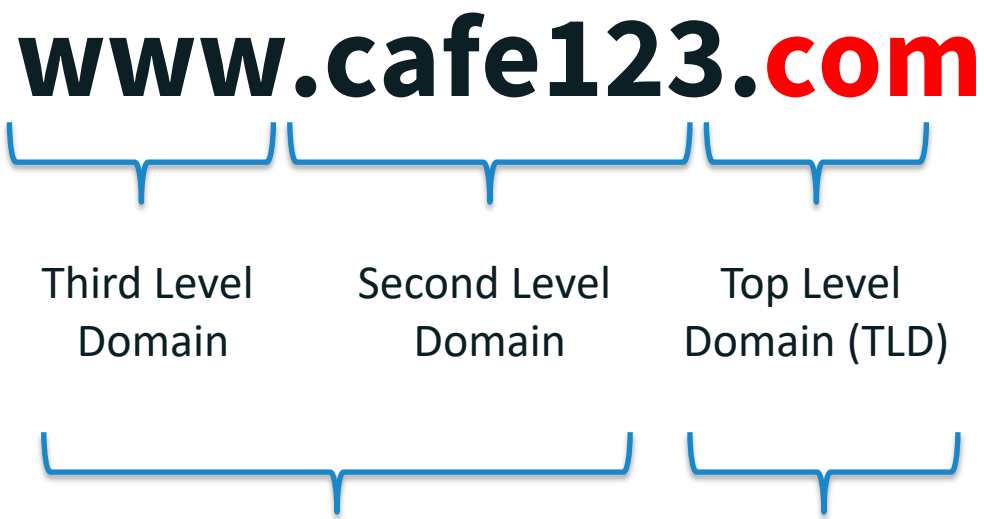
1010 – 1030	Tea and Coffee Break	
1030 – 1200	*Scope of Work* - Languages - Unicode and character properties - IDNA2008 - Maximal Starting Repertoire *Analysis of Code Point Repertoire* - Inclusion and Exclusion	Pitinan Kooarmornpatana
1200 – 1330	Lunch Break	
1330 – 1400	*Analysis of Code Point Repertoire* - Categorization of Code Points	Pitinan Kooarmornpatana
1400 – 1500	Exercise on Code Point Repertoire	ALL
1500 – 1520	Tea and Coffee Break	
1520 – 1600	*Analysis of Code Point Variants* - Within-Script Variants - Cross-Script Variants	Pitinan Kooarmornpatana
1600 – 1700	Exercise on Code Point Variant	ALL
1700 – 1730	Day-1 Summary, Q & A	Thin Zar Phyo

Agenda

Day 2 – Sunday 17 June, 2018		
0900 – 0910	Recap on Day1 and Agenda	Thin Zar Phyo
0910 – 0920	Introduction to Whole Label Evaluation Rules	Pitinan Kooarmornpatana
0930 – 0945	Examples from Devanagari Script	Pitinan Kooarmornpatana
0945 – 1000	Exercise on WLE Rules	ALL
1000 – 1020	Tea and Coffee Break	
1030 – 1100	*Output of Generation Panel* - LGR proposal - LGR XML, LGR Toolset - Test Labels	Pitinan Kooarmornpatana
1100 – 1200	***Myanmar Generation Panel Formation*** - Background - Membership - Work Plan	Thin Zar Phyo
1200 – 1330	Lunch Break	
1330 – 1500	Finalize Myanmar Generation Panel Formation Proposal	Myanmar GP
1500 – 1530	Tea and Coffee Break	
1530 – 1645	Planning the next step	Myanmar GP
1645 – 1730	Meeting Summary, Q & A	Myanmar GP

Overview of IDN Program

ASCII Domain Name Label



Forming ASCII Labels

Use LDH

- **L**etters [a-z]
- **D**igits [0-9]
- **H**yphen (LDH)

Label length = 63

Other constraints (e.g. on hyphen)

Forming ASCII Labels

Use only Letters

- Letters [a-z]

Label length = 63

Internationalized Domain Name (IDN) Labels

Syntax of IDN Labels

Valid U-Label: Unicode code points as constrained by IDNA 2008

Valid A-Label - "xn--" followed by punycode of U-Label of length 59

Syntax of IDN Labels

Valid U-Label, further constrained by the "letter" principle for TLDs

Valid A-Label

“Same” or Different Domain Labels?

- Example of within-script variant labels (Arabic script)

  شبـكـتة (06C3 06A9 0628 0634)
  شبـكـة (0629 06A9 0628 0634)
  شبـكـة (0629 0643 0628 0634)

- Example of within-script variant labels (Simplified Chinese and Traditional Chinese)

  名称 (540D 79F0)
  名稱 (540D 7A31)

- Example of cross-script variant label (Latin script and Cyrillic script)

  epic (0065 0070 0069 0063)
  epic (0435 0440 0456 0441)

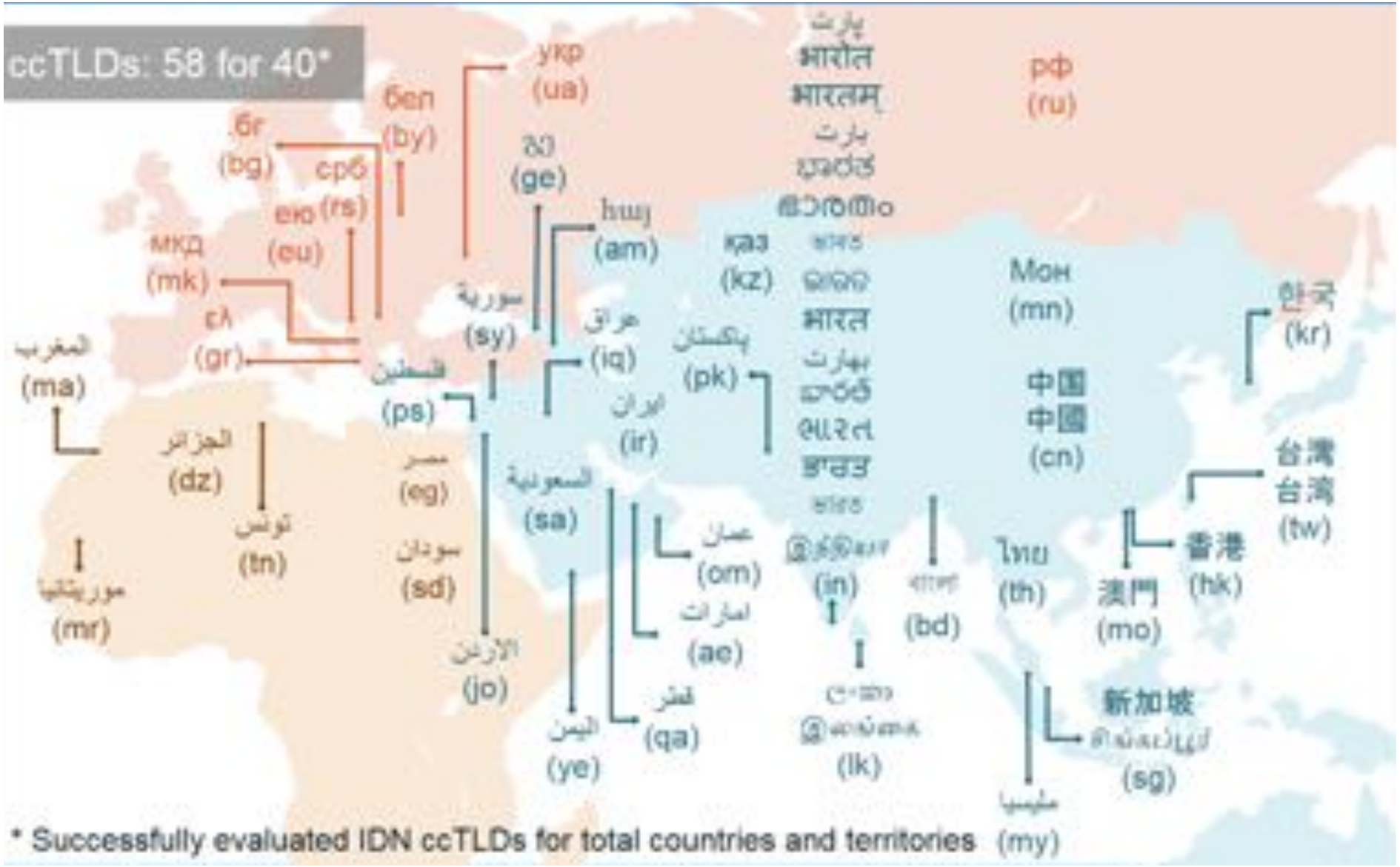
Complex Scripts Require Label Level Rules

- Allow only specific sequences in the script, otherwise could create unpredictable rendering issues for labels
  - Rendering issues can cause confusion in reading labels and potentially security issues for end users
- Prevent labels which should not be allowed or usable
- But should not try to encode language specific spelling rules to allow for innovative labels

IDN Program Objectives

Enable deployment of domain names in the local languages and scripts used by the communities globally in a secure and stable manner

IDN Country Code Top-Level Domains



Root Zone Label Generation Rule Project: Scope of Work

What is the Overall Goal?

- Goal is to create a mnemonic system for use in the Domain Name System (DNS)
  - A mechanism to remember IP address
  - Must remain secure and stable in use – if DNS is confusing to users, then the motivation is not met
  - Not required to completely cover a language or a script
  - May not form labels which are words in a language
    - Not restricted to “correct” spellings
    - May not carry a meaning in the “lexical” sense

Label Generation Rules for the Root Zone

- For the Root Zone, single “table” containing data for all scripts
  - As it is a shared resource, must be conservative
  - Must be stable and secure
  - Must be based on inclusion based analysis
- For each script or writing system:
  - Which code points are valid for use?
  - Are any of these code points variants of each other?
  - Are the any additional constraints on the labels?
- Final work of the Generation Panel:
  - LGR Proposal (XML)
  - LGR Supporting Document (PDF)
  - Test Labels File (TXT)

Root Zone LGR Procedure

Generation Panels

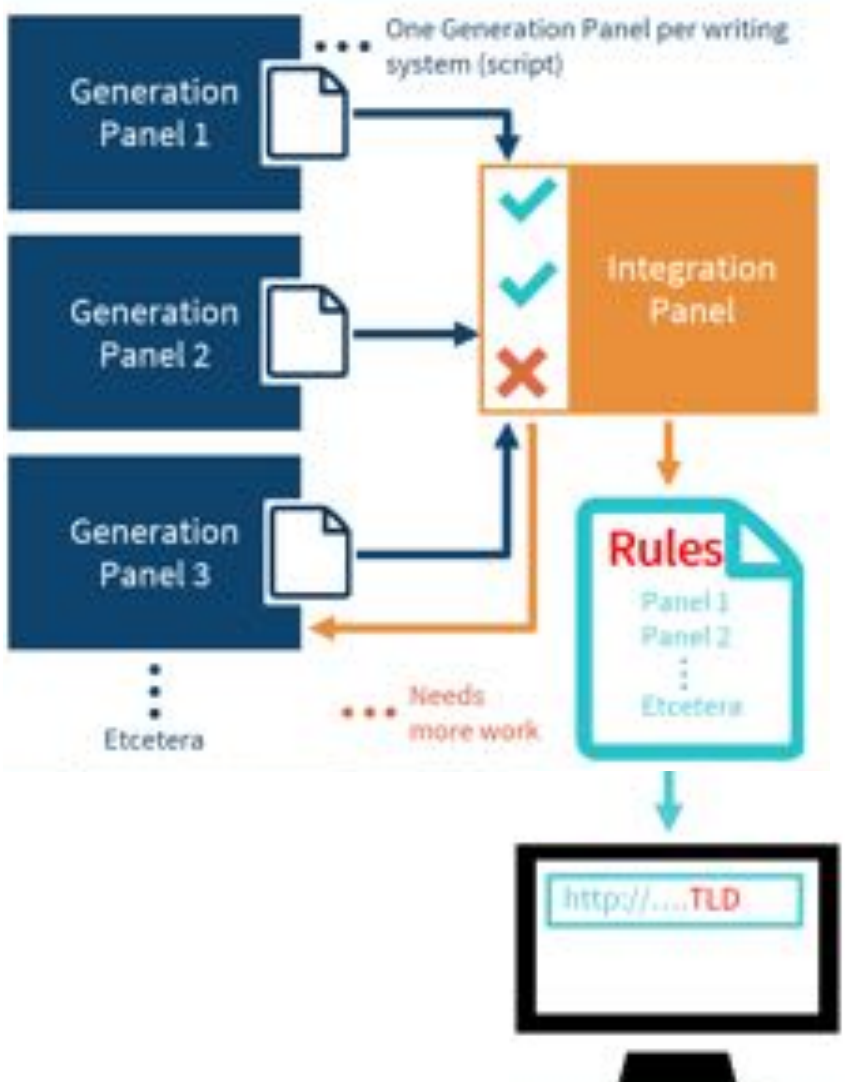
- Generate proposals for script specific LGRs, based on community expertise and requirements

Integration Panel

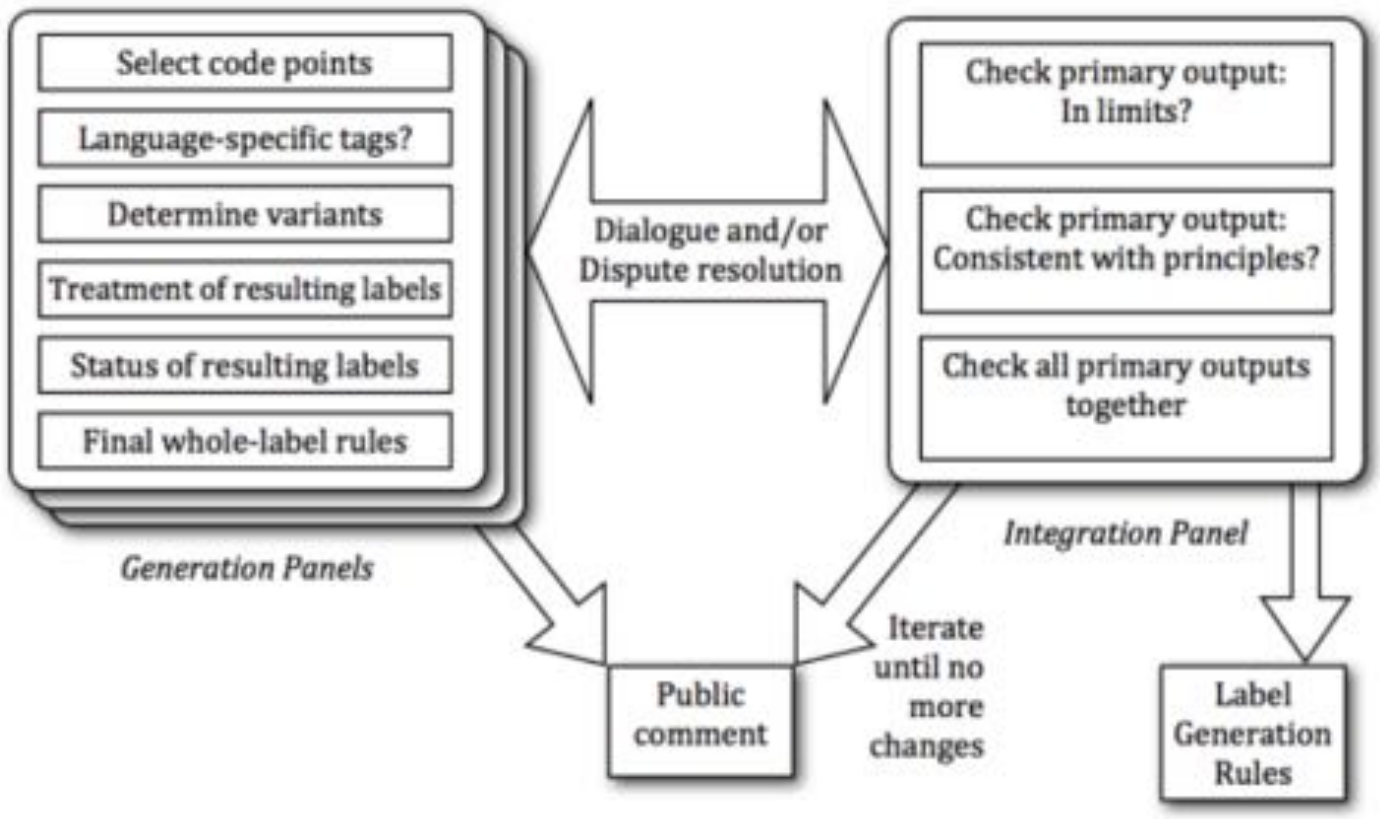
- Integrates them into common Root Zone LGR while minimizing the risk to Root Zone as shared resource

Label Generation Rules (LGR)

- Which labels are permissible
- Which variant labels exist
- Which variant labels may be allocated

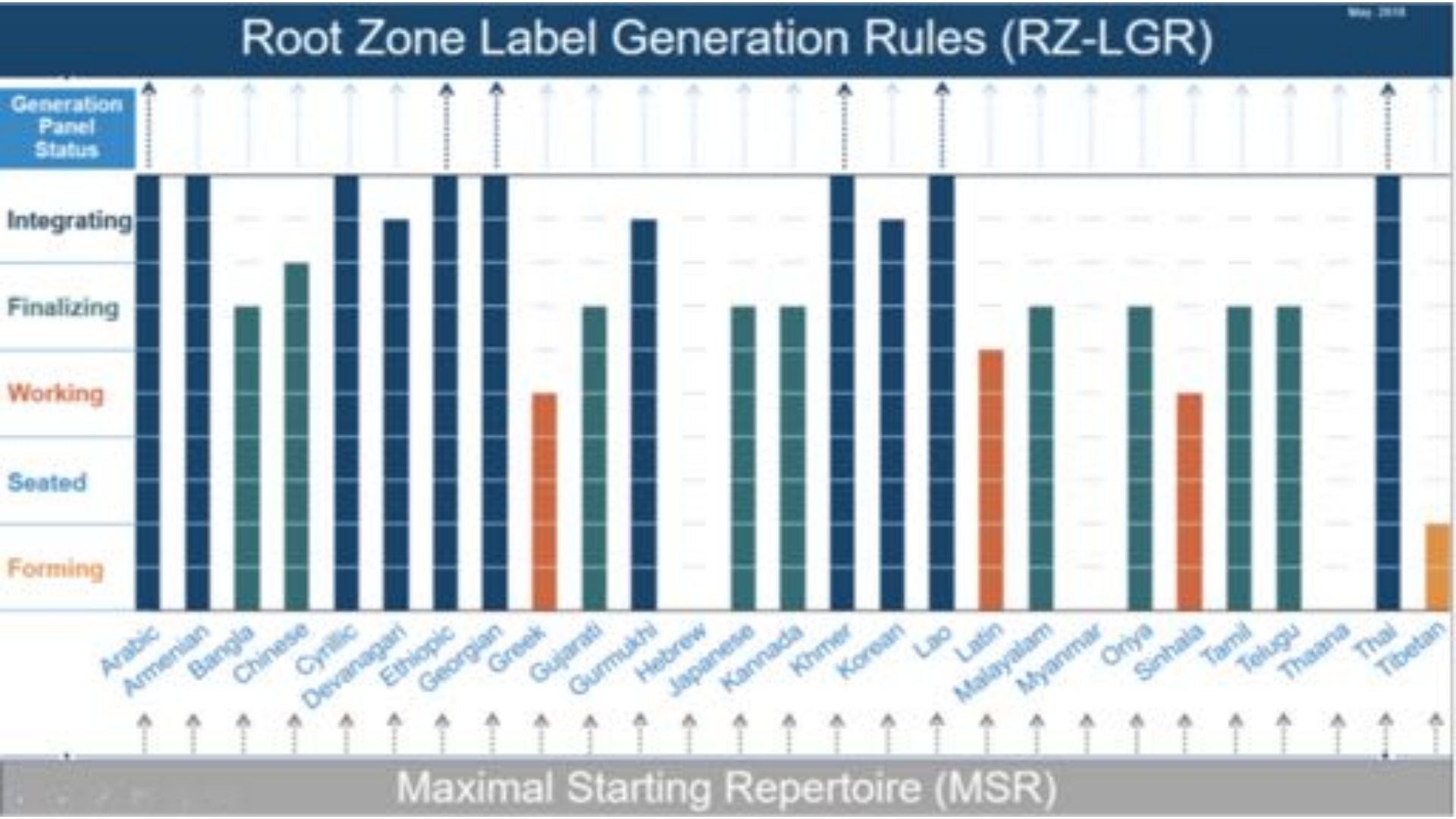


High Level Process



Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels
https://www.icann.org/en/system/files/files/draft-lgr-procedure-20mar13-en.pdf

Status of Generation Panels



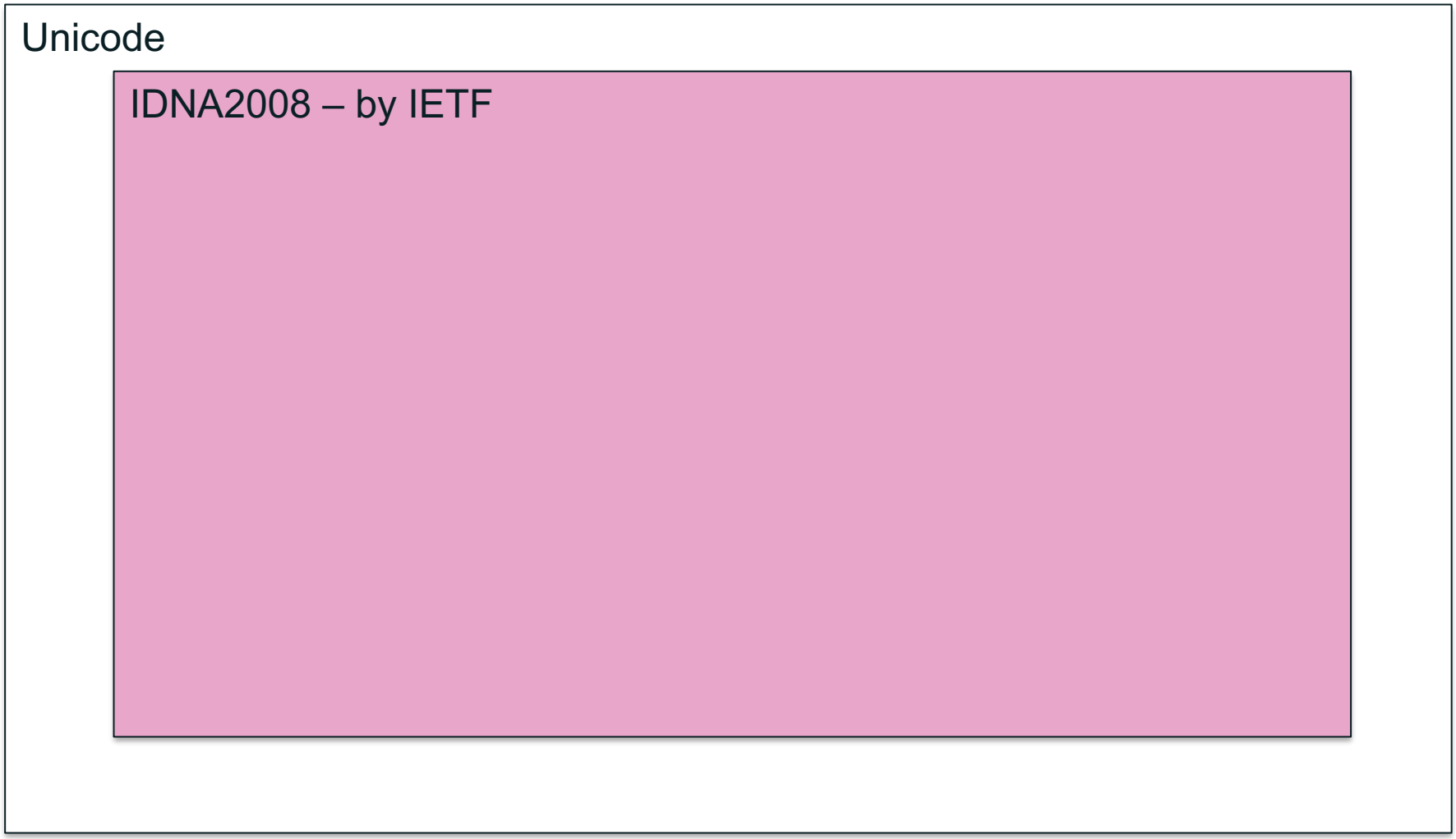
LGR for the Root Zone

Unicode

...

...

LGR for the Root Zone



LGR for the Root Zone

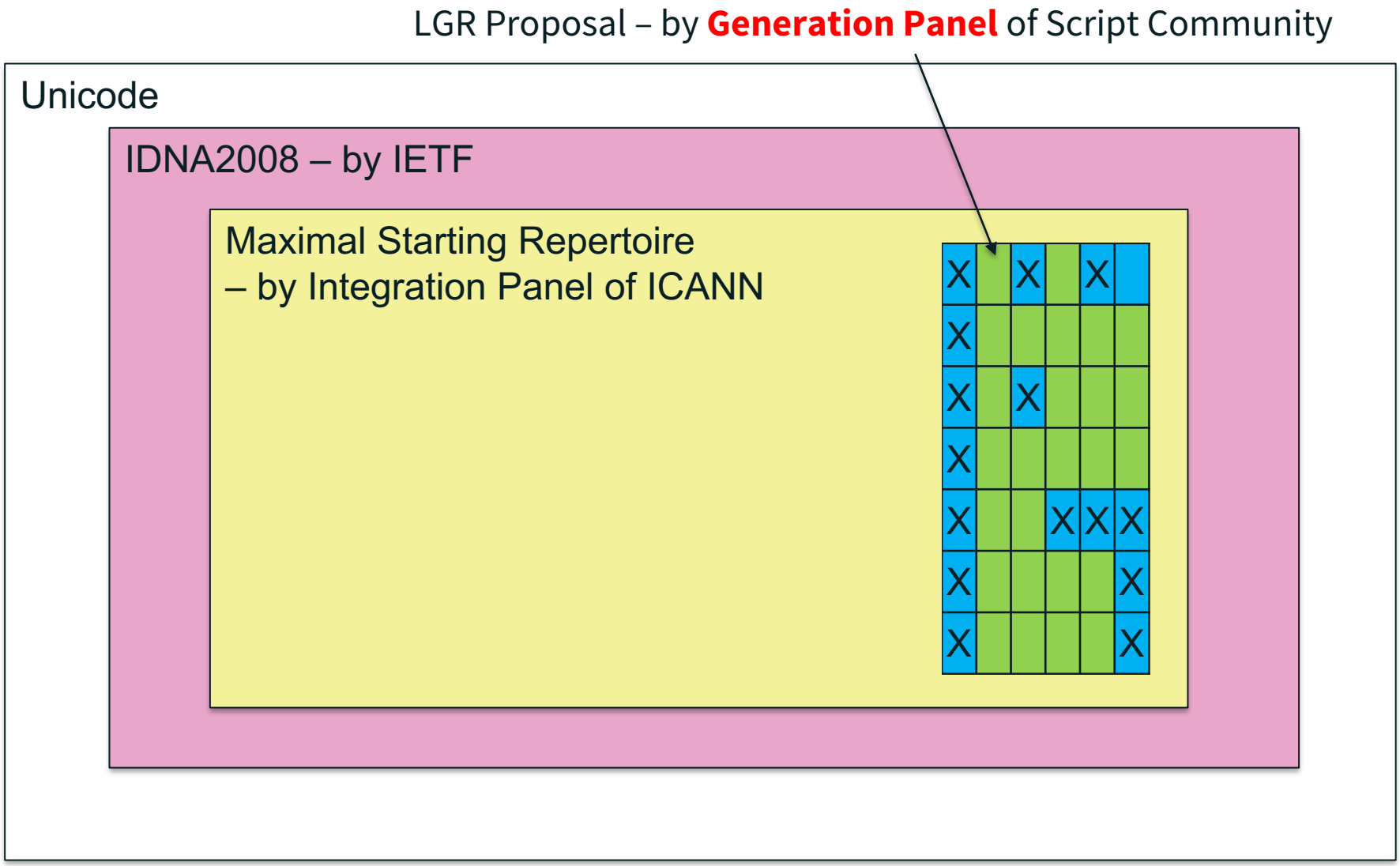
Unicode

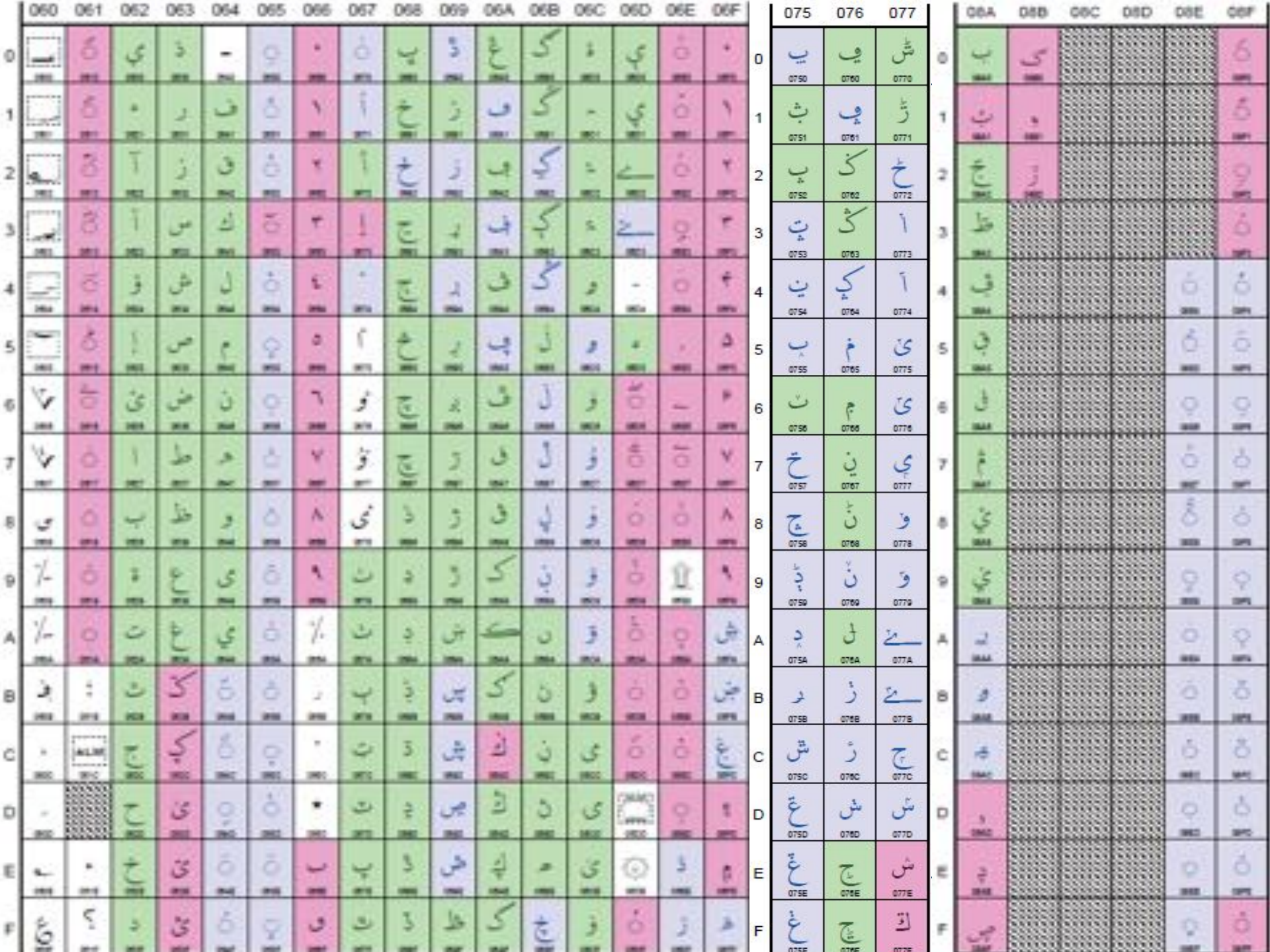
IDNA2008 – by IETF

Maximal Starting Repertoire
– by Integration Panel of ICANN

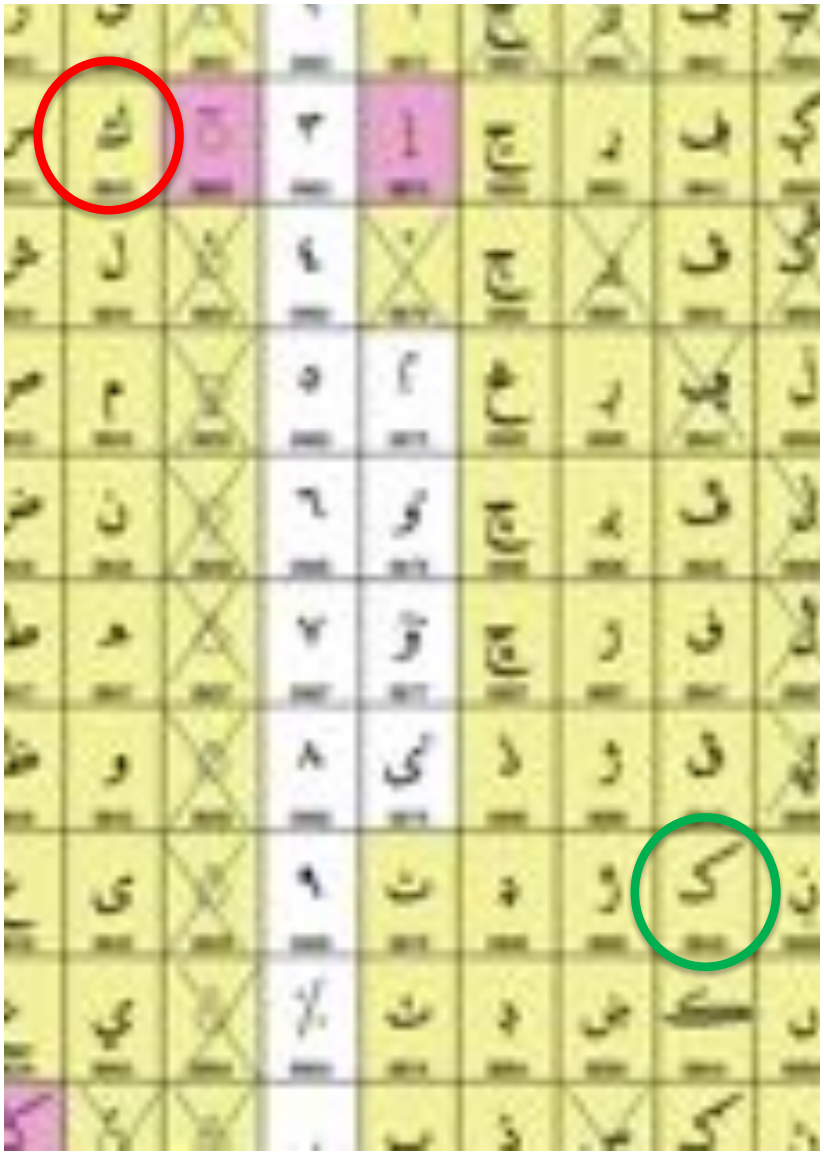
MSR-3:
https://www.icann.org/news/announcement-2-2018-03-29-en

LGR for the Root Zone




060 061 062 063 064 065 066 067 068 069 06A 06B 06C 06D 06E 06F
075 076 077
06A 06B 06C 06D 06E 06F

Label Generation Rules (LGR)



- Valid code points
- Variants code points

كابل

كابل

- Label constraints
  - Cannot mix ک and ك in a label

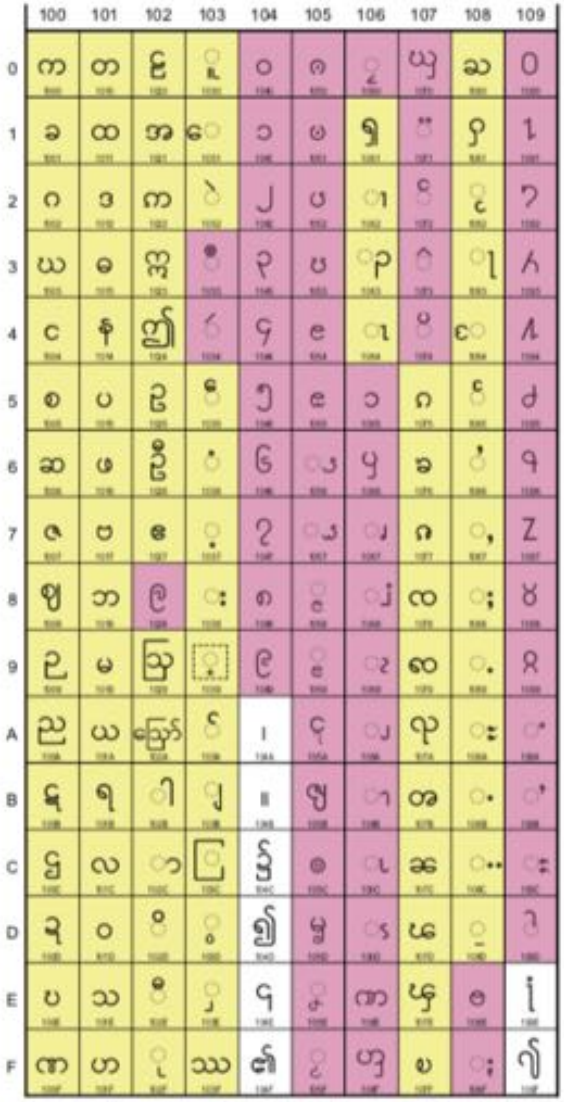
✓ كلكته

✓ كلكته

X كلكته

X كلكته

LGR

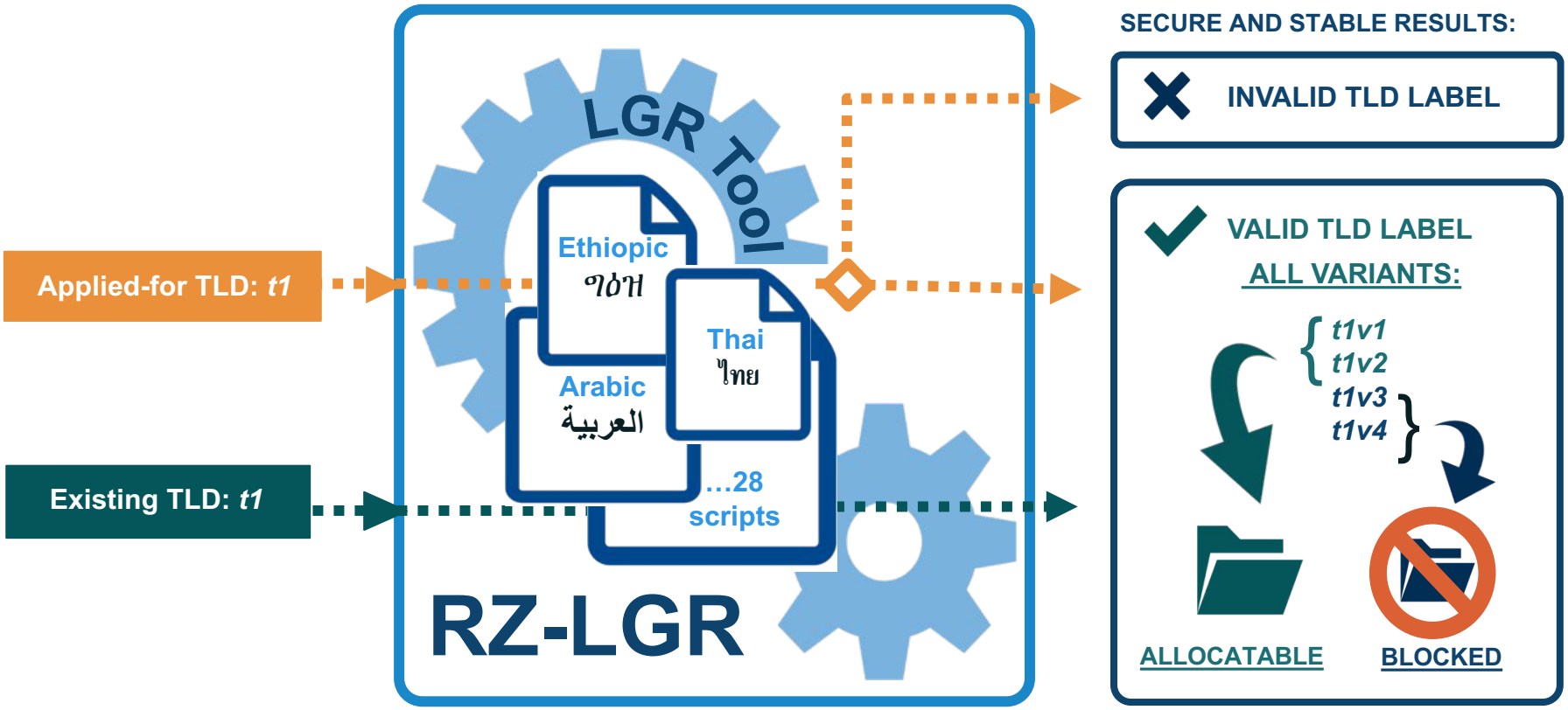


1. Which code points must be included in the Root Zone
   - Are exclusions from MSR (pink) correct?
   - What must be included in LGR?
     - *"everyday, general purpose [use ...] in a stable and widespread manner"*

2. Are there any variant code points
   - Two code points when replaced produce labels considered confusingly similar by an end-user

3. Are there any label-level constraints
   - Well-formedness of a cluster?
   - Constraints on initial or final position in a label?
   - Other?

RZ-LGR for IDN Variant Implementation



LGR Specification

- Label Generation Rulesets (LGRs) used to generate domain name labels, as specified in RFC 7940

```
<?xml version="1.0"?>
<lgr xmlns="urn:ietf:params:xml:ns:lgr-1.0">
    <meta>
        ...
    </meta>
    <data>
        ...
    </data>
    <rules>
        ...
    </rules>
</lgr>
```

Repertoire

What is the Goal?

- Goal is to create a mnemonic system for use in the Domain Name System (DNS)
  - A mechanism to remember IP address
  - Must remain secure and stable in use – if DNS is confusing to users, then the motivation is not met
  - Not required to completely cover a language or a script
  - May not form labels which are words in a language
    - Not restricted to “correct” spellings
    - May not carry a meaning in the “lexical” sense

Starting Point – RFC 6912

Principles

1. Longevity – stable across Unicode versions
2. Least Astonishment– take into account the population using a code point
3. Contextual Safety – sensitive to ways in which code point may be used in malicious ways
4. Conservatism – any code point inclusion decision is as conservative as practicable

Starting Point – RFC 6912

Principles

5. Inclusion – default is excluded, then add code point which is safe based on usability and confusability
6. Simplicity – rules determining use should be simple to understand
7. Predictability – rules determining whether a code point is included are predictable for others to reach the same conclusion
8. Stability – if permitted, taking it out very hard

Starting Point – RFC 6912

Principles

5. Letter – Code point “will be alphabetic” in RFC 1123. Same principle so exclude code points not normally used to write words or used for purposes other than writing words

Questions to Ask

1. Is it contained in the Maximal Starting Repertoire?
2. Is it used with the script defined in the scope of the GP
3. Is it suitable in identifiers?
    a. Is it in widespread modern use?
    b. Is it not technical / religious / limited use only?
    c. Is it not really a punctuation / symbol?
    d. Is it really necessary for representing identifiers?
4. Is the Unicode encoding of the code point stable?
    a. Are there any rendering issues?

Questions to Ask

5. What are the DNS security & stability concerns? rendering issue, homoglyph of non-PVALID code points?
6. How accessible would a TLD containing that code point be?
    a. Are there input/keyboard concerns?
7. What are the risks if the code point is not included?
8. What are the risks if it is?
9. Is it in tension with any of the Principles in any way?
10. Does it always appear in a fixed sequence?

“everyday, general purpose [use ...]”

Level	Label	Description
0	International	The language is widely used between nations in trade, knowledge exchange, and international policy.
1	National	The language is used in education, work, mass media, and government at the national level.
2	Provincial	The language is used in education, work, mass media, and government within major administrative subdivisions of a nation.
3	Wider Communication	The language is used in work and mass media without official status to transcend language differences across a region.
4	Educational	The language is in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education.
5	Developing	The language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable.

https://www.ethnologue.com/about/language-status

“everyday, general purpose [use ...]”

6a	Vigorous	The language is used for face-to-face communication by all generations and the situation is sustainable.
6b	Threatened	The language is used for face-to-face communication within all generations, but it is losing users.
7	Shifting	The child-bearing generation can use the language among themselves, but it is not being transmitted to children.
8a	Moribund	The only remaining active users of the language are members of the grandparent generation and older.
8b	Nearly Extinct	The only remaining users of the language are members of the grandparent generation or older who have little opportunity to use the language.
9	Dormant	The language serves as a reminder of heritage identity for an ethnic community, but no one has more than symbolic proficiency.
10	Extinct	The language is no longer used and no one retains a sense of ethnic identity associated with the language.

https://www.ethnologue.com/about/language-status

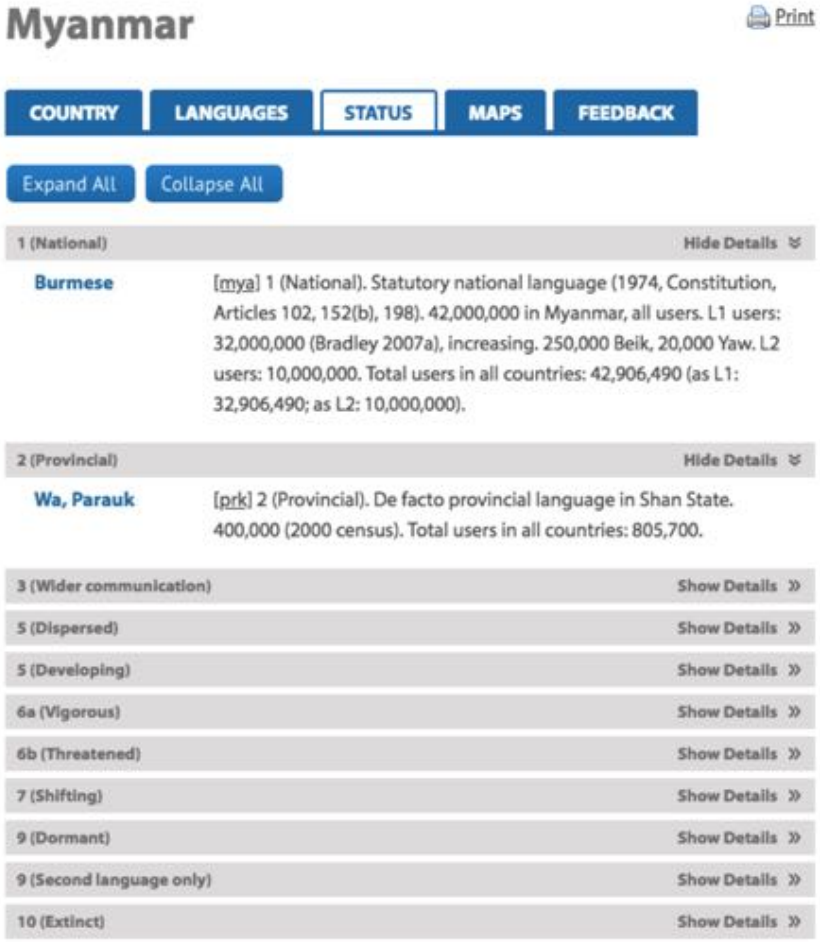
How to Document the Repertoire

- Document general *but relevant* information
  a) History of script
  b) Script characteristics
  c) Languages using the script – standard name, ISO 639 code, name in local script, places language is spoken, other relevant information (e.g. EGIDS no.)
  d) Criteria of language included in analysis (and excluded from analysis)
  e) Types of code points – which types are included and which code points are excluded
  f) Table of code points – with evidence/reference of use for each code point and any additional relevant information

Source of Information - Languages

- National governmental sources
- www.ethnologue.com website
- www.omniglot.com website
- Published research and books
- Field research
- Others?

Source of Information - Languages



Myanmar

Print

COUNTRY | LANGUAGES | STATUS | MAPS | FEEDBACK

Expand All | Collapse All

1 (National) — Hide Details

Burmese — [mya] 1 (National). Statutory national language (1974, Constitution, Articles 102, 152(b), 198). 42,000,000 in Myanmar, all users. L1 users: 32,000,000 (Bradley 2007a), increasing. 250,000 Beik, 20,000 Yaw. L2 users: 10,000,000. Total users in all countries: 42,906,490 (as L1: 32,906,490; as L2: 10,000,000).

2 (Provincial) — Hide Details

Wa, Parauk — [prk] 2 (Provincial). De facto provincial language in Shan State. 400,000 (2000 census). Total users in all countries: 805,700.

3 (Wider communication) — Show Details

5 (Dispersed) — Show Details

5 (Developing) — Show Details

6a (Vigorous) — Show Details

6b (Threatened) — Show Details

7 (Shifting) — Show Details

9 (Dormant) — Show Details

9 (Second language only) — Show Details

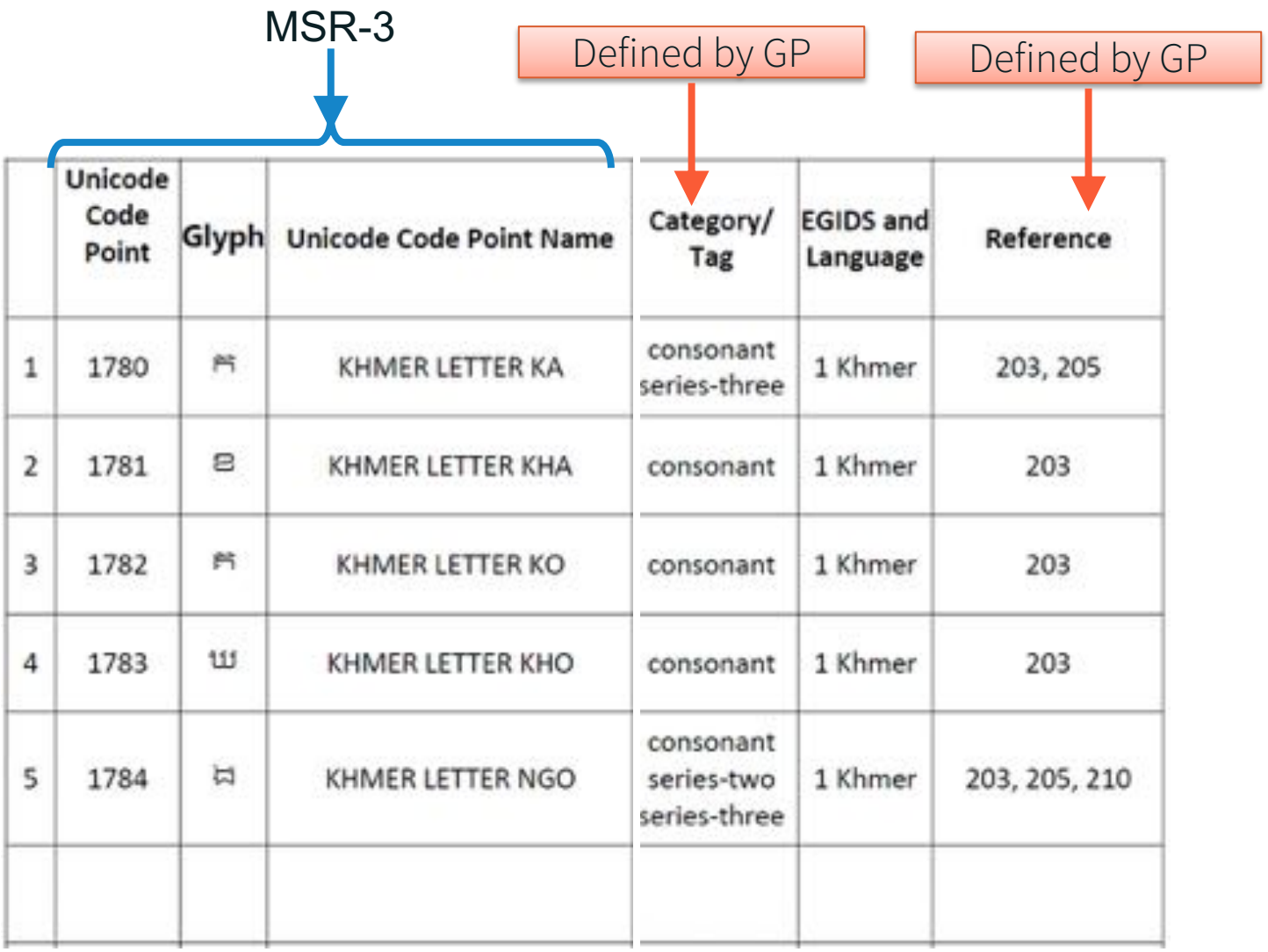
10 (Extinct) — Show Details

- Ethnologue https://www.ethnologue.com/country/MM/status

Sources of Information - Repertoire

- References which could be used to demonstrate “everyday, general purpose [use …]”
  a) National standard published by the government
  b) Books published by Ministry of Education, e.g. for primary school
  c) Common publications, e.g. newspapers
  d) Other?

Example



	Unicode Code Point	Glyph	Unicode Code Point Name	Category/ Tag	EGIDS and Language	Reference
1	1780	ក	KHMER LETTER KA	consonant series-three	1 Khmer	203, 205
2	1781	ខ	KHMER LETTER KHA	consonant	1 Khmer	203
3	1782	គ	KHMER LETTER KO	consonant	1 Khmer	203
4	1783	ឃ	KHMER LETTER KHO	consonant	1 Khmer	203
5	1784	ង	KHMER LETTER NGO	consonant series-two series-three	1 Khmer	203, 205, 210

Exercise

https://docs.google.com/document/d/1fmi8en59uHXuncjpb5fL2B5q04wjt3sfHhPSLFaI_SI/edit#

Variants

What is the Goal?

- Successfully defining variant rules for an LGR is not trivial
- Code point or code point sequences causing two (or more) labels functionally “the same” in a script
- Make the mnemonic system to minimize user confusion
- Conservatism requires
  - maximizing “blocked” variants
  - minimize “allocatable” variants

Questions to Ask

1. Would a reasonable person with native knowledge of the script consider a pair of code points interchangeable?
2. Would such a person be unable to determine which of these interchangeable code points was used by appearance?
3. Is there an alternative representation?
4. What should the disposition of any defined variants be?
5. Should any of the variants of this code point be contingent on context?
6. Is each set of code point variant mappings symmetric?

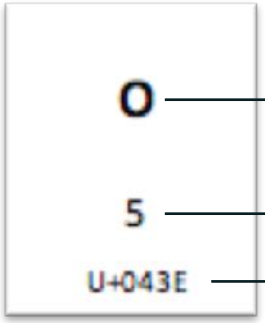
Questions to Ask

7. Is each set of code point variant mappings transitive?
8. Are any variants contemplated that are in tension with any of the Principles?
9. Are the variants designed so that they lead to the minimal required number of allocatable variant labels?
10. Are the variants designed so that, in doubtful cases, they block potential variant labels?
1.

Variant Relationships and Types

- Variants are symmetric
  - A = B ⇒ B = A
- Variants are transitive
  - A = B and B = C ⇒ A = C
- Variant code points can be of two types
  - Allocatable
  - Blocked
- The types are directional
- Label disposition calculated based on types of individual code points
  - A single blocked type causes the whole label to be blocked

Example



If the draw = 5 →

VARIANT BINGO				
ф 37 U+03C6	p 42 U+0070	o 5 U+043E	ш 35 U+0448	ب 12 U+0752
उ 25 U+0A24	ک 11 U+06A9	o 5 U+03BF	ف 10 U+06A2	○ 33 U+0B20
○ 33 U+0D20	ш 35 U+0448	☆	ൽ ീ 32 U+0B88 U+0BCD U+0BB0 U+0BC0	ਇ 24 U+0A07
从 17 U+4E1B	p 42 U+0070	p 42 U+0440	औ 22 U+0914	ㄐㅏ 18 U+535F
ڄ 9 U+0683	x 43 U+0445	= 28 U+0957	उ 25 U+0909	o 5 U+0585

Exercise

https://docs.google.com/document/d/1fmi8en59uHXuncjpb5fL2B5q04wjt3sfHhPSLFal_SI/edit#

Whole Label Evaluation (WLE) Rules

Goal

- Goal is to reduce label space
  - Preventing labels which should not be possible for various reasons
    - Not licensed by the script (but not spelling rules)
    - Cause security issues
    - Cause usability constraints
    - Other?
  - Reducing allocatable label by making them blocked in certain cases
  - Put in contextual contexts for code points or their sequences

Example

- Cannot mix Persian Kaf and Arabic Kaf
- Combining vowel mark must follow a consonant in Lao script
- Subjoining consonant must follow a consonant in Khmer script
- A label cannot start with a combining mark

LGR Proposal (XML)

LGR Specification

- Label Generation Rulesets (LGRs) used to generate domain name labels, as specified in RFC 7940

```
<?xml version="1.0"?>
<lgr xmlns="urn:ietf:params:xml:ns:lgr-1.0">
    <meta>
        ...
    </meta>
    <data>
        ...
    </data>
    <rules>
        ...
    </rules>
</lgr>
```

Test Labels File (TXT)

What Should a Test Labels File Have?

- Have a list of labels
  a. At least 50 valid labels, possibly more, also addressing the following:
     i. Labels should cover all languages shortlisted for inclusion in the proposal for that script (EGIDS 1-4)
     ii. Labels should cover all the code points in the repertoire
  b. A reasonable list of invalid labels containing some of the out-of-repertoire code points

- Have a list of labels which have variant labels (if the script proposal has variants listed), following by a list of all within-script and cross-script variant labels for that label. A reasonable coverage of variant code points is suggested

- Have a few labels for each rule
  a. Have positive examples, creating valid labels
  b. Have negative examples, with invalid labels as they should fail that specific rule – in some cases you may have to make us some nonsense labels just for that purpose
  c. In case there are multiple ways a rule can fail, please suggest a few examples for each of those cases

Questions to Ask about Test Labels File

1. Does it cover all inclusion code points?
2. Does it cover exclusion code points?
3. Does it cover all in-script variant sets?
4. Does it cover all cross-script variant sets?
5. Does it cover all rules for valid cases?
6. Does it cover all rules for invalid cases?

Example of the Test Labels File

- Code point coverage test

```
# Labels-CyrillicScript-20180403
# valid label in repertoire
пойти
дом
тоже
хороший
:

# invalid label out of repertoire
ทดสอบ
```

- Variant set test

```
#valid labels with corresponding list of variants
# сидеть - 0 allocatable variants, 3 blocked variants
сидеть
сидеть
сидеть
сидеть

# мён - 0 allocatable variants, 1 blocked variants
мён
мён
```

Example of the Test Labels File

- WLE test

```
# Pass - Context of NIKAHIT SIGN (U+17C6)
កុំ
សន្សំ
បង្គំ
យំ
ចាំ
ទាំង
ខ្ញុំ
# Fail - Context of REAHMUK SIGN (U+17C7)
ក្រឹះ
ឯកភ
ព្រះសាន
ឥះដ្ឋ
ឫះស្សី
```

More examples: https://www.icann.org/resources/pages/lgr-proposals-2015-12-01-en

Engage with the IDN Program

Thank You and Questions

Visit us at **icann.org/idn**
Email: IDNProgram@icann.org

 @icann

 facebook.com/icannorg

 youtube.com/icannnews

 flickr.com/icann

 linkedin/company/icann

 slideshare/icannpresentations

 soundcloud/icann