



Neo-Brahmi Scripts LGR for the Root Zone

Sarmad Hussain | Director IDN Programs | 24-26 May 2017

Agenda

1

Background
of TLD
Program

2

Repertoire

3

Variants

4

WLE Rules

5

XML
Formulation
and LGR Tool

6

Conclusions
and Next
Steps



Background of TLD Program

ASCII Domain Name Label

www.cafe123.com



Third Level
Domain

Second Level
Domain

Top Level
Domain (TLD)

Top Level Domains (TLDs)

- ⦿ Country Code TLDs (ccTLDs)
 - ⦿ .sg, .cn, .kh, .la, .mm, .th, .ca, ...
 - ⦿ Two letter [a..z] codes, reserved for countries and territories by ISO 3166 standard
- ⦿ Generic TLDs (gTLDs)
 - ⦿ .com, .org, .net, .edu, ... - organizations
 - ⦿ New gTLDs – 1930 applications in 2012

Domain Stakeholders

- ⦿ ICANN
- ⦿ Registry
- ⦿ Registrar
- ⦿ Reseller
- ⦿ Registrant
- ⦿ End-User

ASCII Domain Name Label

www.cafe123.com

Third Level
Domain

Second Level
Domain

Top Level
Domain (TLD)

Forming ASCII Labels

Use LDH

- Letters [a-z]
- Digits [0-9]
- Hyphen (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

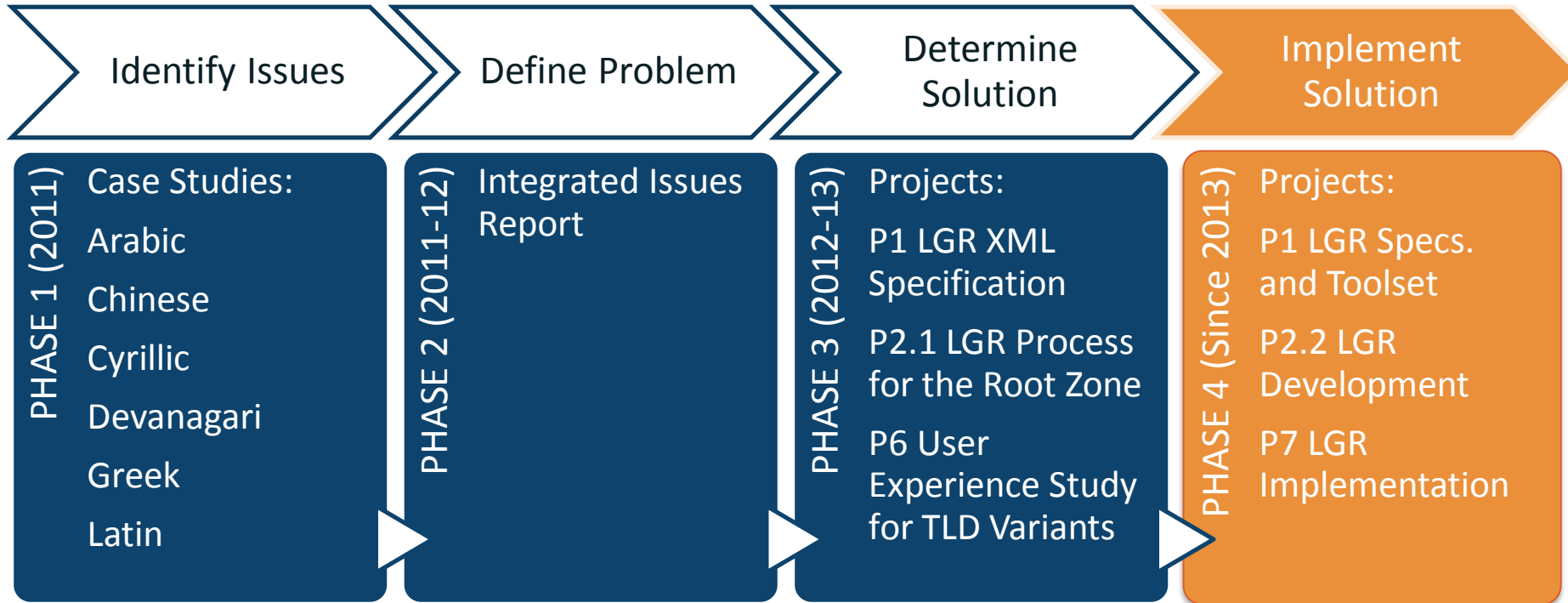
Valid A-Label - “xn--” followed by punycode of U-Label of length 59

Syntax of IDN Labels

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

Valid A-Label

IDN TLD Program



Community agreed to define a Label Generation Rules (LGR)

Reports and documentation of all completed projects available at:
<https://www.icann.org/resources/pages/reports-2013-04-03-en>

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 there any additional constraints on the labels?

LGR for the Root Zone

Unicode

	000	001	002	003	004	005	006	007
0	NUL	DLE	SP	0	@	P	`	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	\$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(8	H	X	h	x
9	HT	EM)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[k	{
C	FF	FS	,	<	L	\	l	
D	CR	GS	-	=	M]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	DEL

• • •

	120	121	122	123	124	125	126	127	128	129	12A	12B
0	ሀ	ሐ	ሠ	ሰ	ቀ	ቐ	በ	ተ	ጎ	ነ	አ	ኮ
1	ሁ	ሑ	ሡ	ሱ	ቁ	ቑ	ቡ	ቱ	ጊ	ጊ	ኦ	
2	ሂ	ሐ	ሢ	ሲ	ቂ	ቊ	ቢ	ቲ	ጋ	ጋ	ኦ	ኩ
3	ሃ	ሐ	ሣ	ሳ	ቃ	ቋ	ባ	ታ	ጌ	ጌ	ኦ	ኩ
4	ሄ	ሐ	ሣ	ሳ	ቃ	ቋ	ባ	ታ	ጌ	ጌ	ኦ	ኩ
5	ህ	ሐ	ሥ	ስ	ቅ	ቍ	ብ	ቲ	ጎ	ጎ	ኦ	ኮ
6	ሆ	ሐ	ሥ	ስ	ቅ	ቍ	ብ	ቲ	ጎ	ጎ	ኦ	ኮ
7	ሀ	ሐ	ሢ	ሲ	ቂ	ቊ	ቢ	ቲ	ጋ	ጋ	ኦ	ኩ
8	ለ	መ	ረ	ሸ	ቄ	ቆ	ሸ	ቆ	ጎ	ኘ	ከ	ኸ
9	ሉ	መ	ሩ	ሹ			ሹ	ቆ	ጎ	ኘ	ከ	ኸ
A	ሊ	ሚ	ሪ	ሸ	ቀ	ቆ	ሸ	ቆ	ጎ	ኘ	ከ	ኸ
B	ላ	ማ	ሪ	ሸ	ቀ	ቆ	ሸ	ቆ	ጎ	ኘ	ከ	ኸ
C	ሌ	ሚ	ሪ	ሸ	ቀ	ቆ	ሸ	ቆ	ጎ	ኘ	ከ	ኸ
D	ል	ም	ር	ሸ	ቀ	ቆ	ሸ	ቆ	ጎ	ኘ	ከ	ኸ
E	ሎ	ም	ሮ	ሾ			ሾ	ቆ	ጎ	ኘ	ከ	ኸ
F	ሊ	ሚ	ሪ	ሸ			ሸ	ቆ	ጎ	ኘ	ከ	ኸ

• • •

LGR for the Root Zone

Unicode

IDNA2008 – by IETF

LGR for the Root Zone

Unicode

IDNA2008

Maximal Starting Repertoire – by Integration Panel of ICANN

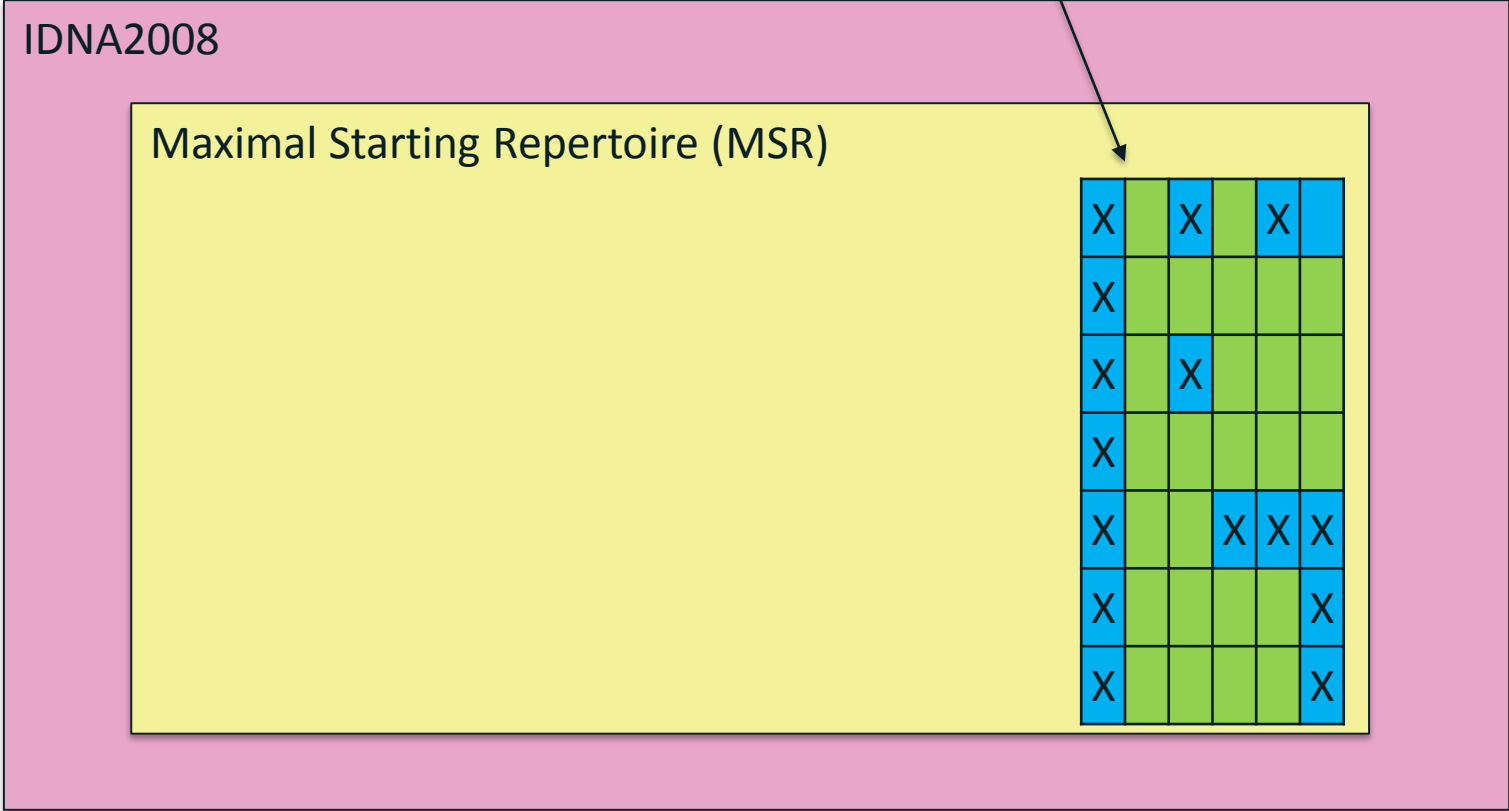
LGR for the Root Zone

LGR Proposal – by **Generation Panel** of Script Community

Unicode

IDNA2008

Maximal Starting Repertoire (MSR)



X		X		X	
X					
X		X			
X					
X			X	X	X
X					X
X					X

	060	061	062	063	064	065	066	067	068	069	06A	06B	06C	06D	06E	06F	075	076	077	08A	08B	08C	08D	08E	08F
0	ا	ب	ي	ذ	-	ر	و	پ	ت	غ	گ	ة	ي	و	و	و	ب	ب	ش	ب	ک				و
1	ا	ب	ء	ر	ف	ا	ا	ا	خ	ز	ف	گ	ه	ي	و	ا	ب	پ	ز	ب	ا				و
2	م	و	آ	ز	ق	و	ا	ا	خ	ز	ب	گ	ا	ا	و	ا	پ	ک	خ	ج	ز				و
3	ب	و	ا	س	ك	ا	ا	ا	ج	ر	ف	گ	ا	ا	و	ا	پ	ک	ا	ظ					و
4	ا	ب	و	ش	ل	و	ا	ا	ج	ر	ف	گ	و	-	و	ا	پ	ک	ا	ق				و	و
5	ا	ب	ا	ص	م	و	ا	ا	خ	ر	پ	ل	و	ه	ر	ا	ب	م	ئ	ق				و	و
6	ا	ب	و	ض	ن	و	ا	و	چ	ر	ق	ل	و	ا	ا	ا	م	م	ئ	ث				و	و
7	ا	ب	ا	ط	ه	و	ا	و	چ	ر	ف	ل	و	ا	ا	ا	پ	ن	ي	م				و	و
8	و	و	ب	ظ	و	و	ا	ئ	ذ	ر	ق	پ	و	و	و	ا	ب	ن	و	ج				و	و
9	ا	و	و	ع	ي	و	ا	ث	د	ر	ك	ن	و	و	ا	ا	ب	ن	و	ب				و	و
A	ا	و	ت	غ	ي	ا	ا	ث	د	ب	ك	ن	و	ا	و	ا	ب	ل	ا	د				و	و
B	ف	؛	ث	ك	و	و	ا	ب	ب	پ	ك	ن	و	و	و	ا	ب	ر	ز	ا				و	و
C	،	ALM	ج	ك	و	و	ا	ب	ت	ت	پ	ن	ي	و	و	ا	ب	ر	ز	ج				و	و
D	ر		ح	ئ	و	و	ا	ب	ت	د	ك	ن	ي	و	و	ا	ب	ر	ش	ش				و	و
E	ا	ء	خ	ئ	و	و	ا	ب	ت	ض	ك	ه	ئ	و	و	ا	ب	ر	ج	ش				و	و
F	غ	؟	د	ئ	و	و	ا	ب	ت	ظ	ك	خ	و	و	و	ا	ب	ر	ج	ك				و	و

Label Generation Rules (LGR)

0632	ك	0652	3	0672	ح	ر	فبا	0682	0692	06A2	06B2
0643	ك	0653	3	0673	ح	ر	فبا	0683	0693	06A3	06B3
0644	ل	0654	4	0674	ح	ر	فبا	0684	0694	06A4	06B4
0645	م	0655	5	0675	ح	ر	فبا	0685	0695	06A5	06B5
0646	ن	0656	6	0676	ح	ر	فبا	0686	0696	06A6	06B6
0647	ه	0657	7	0677	ح	ر	فبا	0687	0697	06A7	06B7
0648	و	0658	8	0678	ح	ر	فبا	0688	0698	06A8	06B8
0649	ي	0659	9	0679	ح	ر	فبا	0689	0699	06A9	06B9
064A	ي	065A	%	067A	ح	ر	فبا	068A	069A	06AA	06BA

- Valid code points
- Variants code points

کابل

کابل

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

کککته ✓

کککته ✓

کککته x

کککته x

MSR and LGR

	090	091	092	093	094	095	096	097
0	ॐ 0900	ऐ 0910	ठ 0920	र 0930	ी 0940	ॐ 0950	ऋ 0960	० 0970
1	ँ 0901	ऑ 0911	ड 0921	र 0931	ु 0941	ं 0951	ृ 0961	· 0971
2	ं 0902	ओ 0912	ढ 0922	ल 0932	ॢ 0942	ॣ 0952	ॡ 0962	ँ 0972
3	ः 0903	ओ 0913	ण 0923	ळ 0933	ॣ 0943	॥ 0953	ॢ 0963	अ 0973
4	ऐ 0904	औ 0914	त 0924	ळ 0934	ॣ 0944	। 0954	। 0964	आ 0974
5	अ 0905	क 0915	थ 0925	व 0935	ॠ 0945	ॡ 0955	॥ 0965	औ 0975
6	आ 0906	ख 0916	द 0926	श 0936	ॢ 0946	ॣ 0956	० 0966	अ 0976
7	इ 0907	ग 0917	घ 0927	ष 0937	ॢ 0947	ॣ 0957	१ 0967	अ 0977
8	ई 0908	घ 0918	न 0928	स 0938	ॢ 0948	ॣ 0958	२ 0968	र 0978
9	उ 0909	ड 0919	न 0929	ह 0939	ॉ 0949	ख 0959	३ 0969	ज़ 0979
A	ऊ 090A	च 091A	प 092A	ं 093A	ो 094A	ग 095A	४ 096A	य 097A
B	ऋ 090B	ॠ 091B	फ 092B	ा 093B	ो 094B	ज़ 095B	५ 096B	ग 097B
C	ॠ 090C	ज 091C	ब 092C	ॢ 093C	ौ 094C	ड़ 095C	६ 096C	ज़ 097C
D	ँ 090D	झ 091D	भ 092D	ॣ 093D	ॢ 094D	ढ़ 095D	७ 096D	१ 097D
E	ऐ 090E	ज 091E	म 092E	ा 093E	ि 094E	फ़ 095E	८ 096E	ड 097E
F	ए 090F	ट 091F	य 092F	ि 093F	ौ 094F	य 095F	९ 096F	ब 097F

- 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”*
- Are there any **variant code points**
 - Two code points when replaced produce labels considered confusingly similar by an end-user
- Are there any **label-level constraints**
 - Well-formedness of a cluster?
 - Constraints on initial or final position in a label?
 - Other?

Root Zone LGR Procedure

Generation Panels

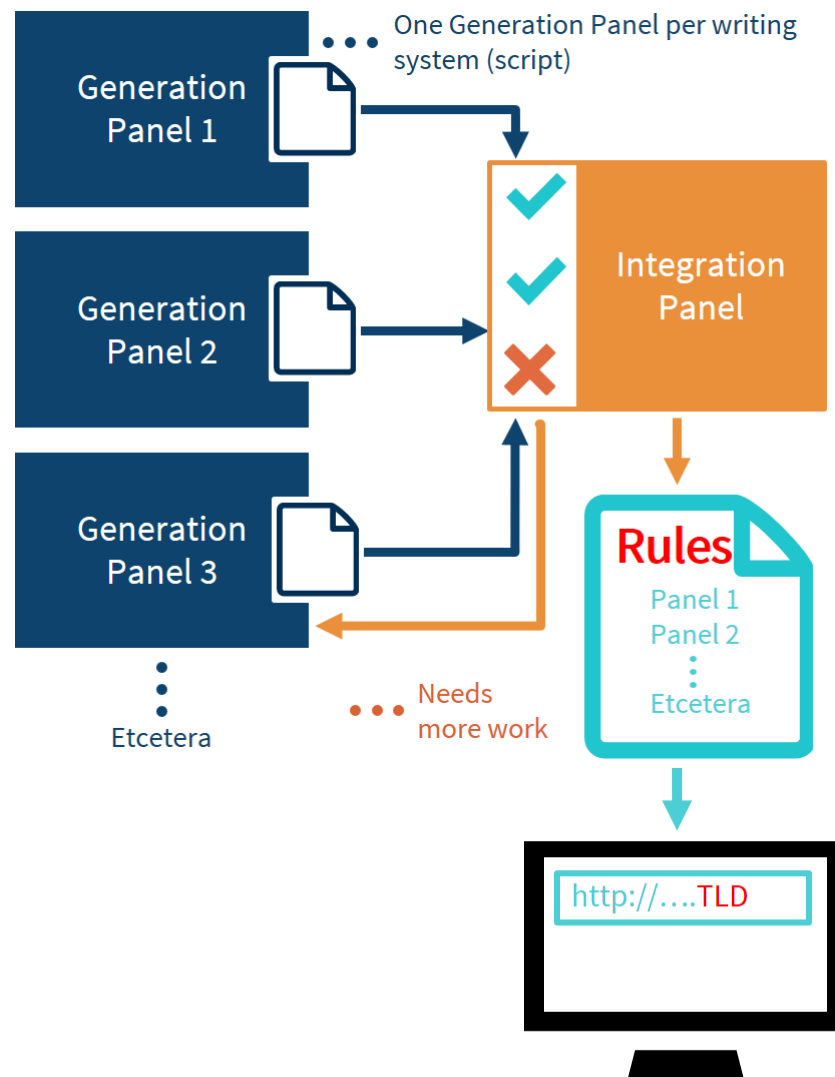
- Generate proposals for script specific LGRs, based on community expertise and requirements

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



LGR Specification

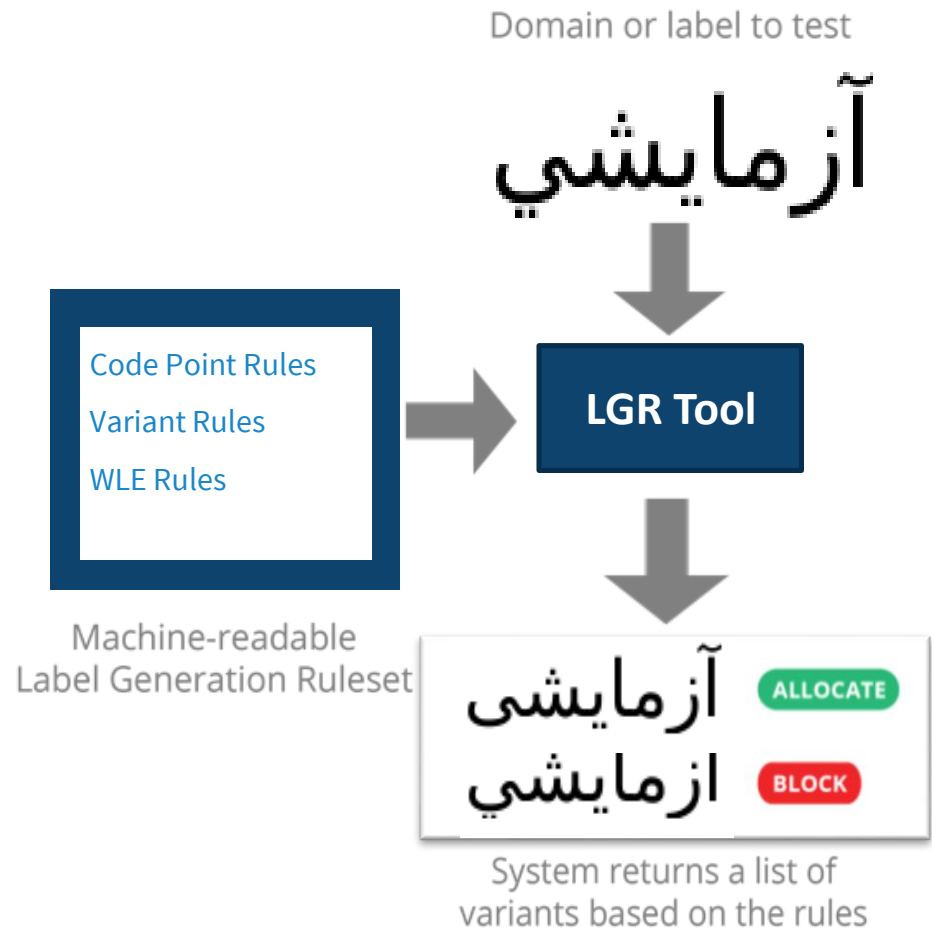
- ◉ Label Generation Rulesets (LGRs) used to generate domain name labels, per [RFC 7940](#)

- ◉ Example: excerpt from MSR-2 XML file

```
...  
<range first-cp="0780" last-cp="07B0" tag="sc:Thaa" ref="3"/>  
<char cp="07B1" tag="sc:Thaa" ref="5"/>  
<char cp="08A0" tag="sc:Arab" ref="12"/>  
<range first-cp="08A2" last-cp="08AC" tag="sc:Arab" ref="12"/>  
<range first-cp="08E4" last-cp="08EF" tag="sc:Arab" ref="12"/>  
<range first-cp="08F4" last-cp="08FE" tag="sc:Arab" ref="12"/>  
<range first-cp="0901" last-cp="0903" tag="sc:Deva" ref="0"/>  
<char cp="0904" tag="sc:Deva" ref="6"/>  
<range first-cp="0905" last-cp="0939" tag="sc:Deva" ref="0"/>  
<range first-cp="093A" last-cp="093B" tag="sc:Deva" ref="11"/>  
<char cp="093C" tag="sc:Deva" ref="0"/>  
<range first-cp="093E" last-cp="094D" tag="sc:Deva" ref="0"/>  
<char cp="094F" tag="sc:Deva" ref="11"/>  
<range first-cp="0956" last-cp="0957" tag="sc:Deva" ref="11"/>  
<char cp="0972" tag="sc:Deva" ref="9"/>  
<range first-cp="0973" last-cp="0977" tag="sc:Deva" ref="11"/>  
<range first-cp="0979" last-cp="097A" tag="sc:Deva" ref="10"/>  
<range first-cp="097B" last-cp="097C" tag="sc:Deva" ref="8"/>  
<range first-cp="097E" last-cp="097F" tag="sc:Deva" ref="8"/>  
<range first-cp="0981" last-cp="0983" tag="sc:Beng" ref="0"/>  
...
```

LGR Toolset (beta)

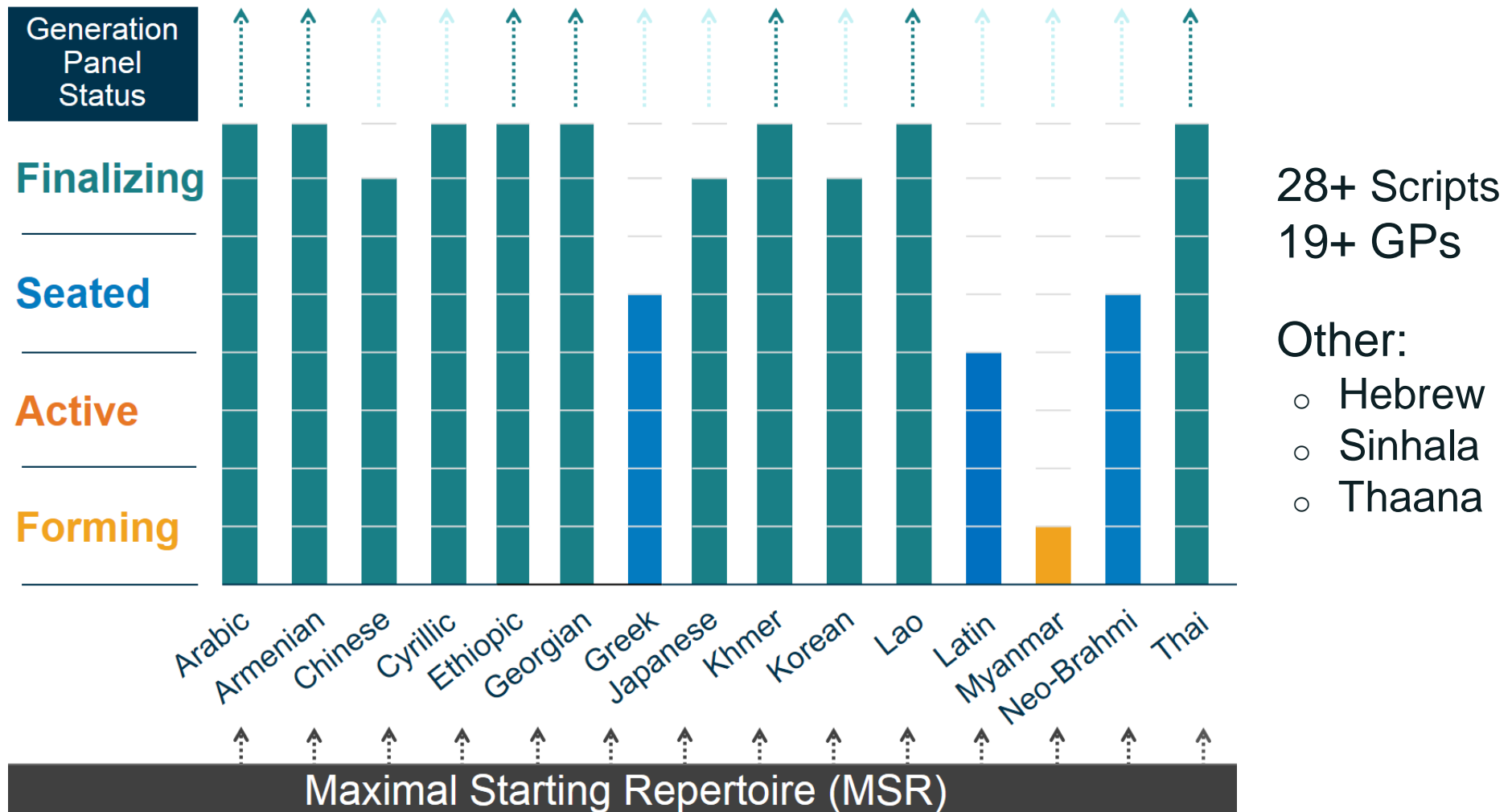
- LGR Toolset allows for the following:
 - Create a LGR
 - Use a LGR to validate label and variants
 - Manage LGRs
- Online beta deployment
 - Visit <https://lgrtool.icann.org/>
- Open source package(s) release
 - Released at github: [lgr-core](#), [lgr-django](#), [munidata](#)
- [User guide](#) available for further details



Root Zone LGR Development Status

Label Generation Rules (LGR)

May 2017



Scope of Neo-Brahmi GP

- On 15 May 2015 the Neo-Brahmi script community submitted to ICANN the [Proposal for Generation Panel for Neo-Brahmī Scripts Label Generation Ruleset for the Root Zone](#), covering Bengali, Devanagari, Gujarati, Gurmukhi, Kannada, Malayalam, Oriya, Tamil and Telugu scripts
 - For each script
 - Cover all languages with widespread, everyday general use (e.g. EGIDS scale less than 4 (or 5?))
 - Cover all geographical regions where the scripts are used
 - Ensure that technical criteria is met along with the socio-linguistic requirements



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

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

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

Principles

9. 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

Sources of Information - Languages

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

Sources of Information - Languages

Nepal



COUNTRY

LANGUAGES

STATUS

MAPS

FEEDBACK

Expand All

Collapse All

1 (National)	Show Details »
3 (Wider communication)	Show Details »
4 (Educational)	Show Details »
5 (Dispersed)	Show Details »
5 (Developing)	Show Details »
6a (Vigorous)	Show Details »
6b (Threatened)	Show Details »
7 (Shifting)	Show Details »
8a (Moribund)	Show Details »
8b (Nearly extinct)	Show Details »
9 (Dormant)	Show Details »
10 (Extinct)	Show Details »

- ⦿ Ethnologue - <https://www.ethnologue.com/country/NP/status>

Sources of Information - Languages

4 (Educational)	Hide Details	5 (Dispersed)	Hide Details
Avadhi	<p>[awa] 4 (Educational). Language of recognized indigenous nationality: Kushwadiya. 547,400 in Nepal, all users. 502,000 (2011 census), increasing. 45,400 (2011 census).</p>	Bengali	<p>[ben] 5 (Dispersed). 23,980 in Nepal, all users. 21,100 (2011 census), increasing. 2,880 (2011 census).</p>
Bhojpuri	<p>[bho] 4 (Educational). 1,740,000 in Nepal, all users. 1,580,000 (2011 census), increasing. 160,000 (2011 census).</p>	Maithili	<p>[mai] 5 (Dispersed). 4,085,000 in Nepal, all users. 3,890,000 (2011 census), increasing. 793,000 Bajjika, 3,090,000 Maithili (2011 census). 195,000 (2011 census).</p>
English	<p>[eng] 4 (Educational). 7,002,030 in Nepal, all users. 2,030 (2011 census). 7,000,000 (Crystal 2003a).</p>	Santhali	<p>[sat] 5 (Dispersed). Language of recognized indigenous nationality: Satar (Santhal). 50,880 in Nepal, all users. 49,900 (2011 census), increasing. 980 (2011 census). Very few monolinguals. Ethnic population: 51,700 (2011 census).</p>
Newar	<p>[new] 4 (Educational). Language of recognized indigenous nationality: Newar. 879,600 in Nepal, all users. 847,000 (2011 census), decreasing. 32,600 (2011 census). Many women are monolingual. Ethnic population: 1,260,000. Includes 1,245,000 Newar and 11,500 Pahari. Total users in all countries: 893,600 (as L1: 861,000; as L2: 32,600).</p>	Urdu	<p>[urd] 5 (Dispersed). 737,800 in Nepal, all users. 692,000 (2011 census). 45,800 (2011 census).</p>
Tamang, Eastern	<p>[taj] 4 (Educational). Language of recognized indigenous nationality: Tamang. 1,213,500 in Nepal, all users. 1,180,000 (2001 census), increasing. Southwestern Tamang: 109,000 (1991 census). Population for all Tamang varieties: 1,350,000 (2011 census). 33,500 (2011 census). In some remote communities, particularly women, children and elderly people are monolingual. Ethnic population: 1,290,000 (2001 census). Total users in all countries: 1,231,000 (as L1: 1,197,500; as L2: 33,500).</p>		
Tibetan	<p>[bod] 4 (Educational). Spoken as a trade language among Bodish groups in Nepal. 4,450 (2011 census), increasing.</p>		

Sources of Information - Languages

Tigrigna



LANGUAGE

FEEDBACK

A language of Ethiopia

ISO 639-3

tir

Alternate Names

Beta Israel, Tigray, Tigrinya

Population

4,320,000 in Ethiopia (2010 UNSD). L2 users: 147,000 in Ethiopia. 2,820,000 monolinguals. Total users in all countries: 7,899,400 (as L1: 7,747,400; as L2: 152,000).

Location

Tigray region border areas; Amhara and Afar regions.

Language Maps

Djibouti, Eritrea and Ethiopia

Language Status

2 (Provincial). Statutory provincial working language in Tigray Region (1994, Constitution, Article 47(3)).

Classification

Afro-Asiatic, Semitic, South, Ethiopian, North

Typology

SOV; noun head final; gender (masculine/feminine); definite article; verb affixes mark person, number, gender of subject; passives; aspect; 33 consonant and 14 vowel phonemes.

Language Use

Also use Amharic [amh]. Used as L2 by Amharic [amh], Kunama [kun], Xamtanga [xan].

Language Development

Literacy rate in L2: 27%. Taught in primary schools. Fully developed. Bible: 1956.

Language Resources

OLAC resources in and about Tigrigna

Writing

Ethiopic script [Ethi], used since 13th or 14th century.

Other Comments

Christian.

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?

Strategies for Documenting the Repertoire

- ⦿ Strategy 1: Code Point Analysis
 - ⦿ For each code point in MSR
 - ⦿ Determine if it is used by one or more languages included
 - one example is sufficient
 - ⦿ Determine if the code point is required for the language(s)
 - ⦿ Document reference and reason for inclusion
- ⦿ Strategy 2: Language Analysis
 - ⦿ For each included language short-listed by GP
 - ⦿ Determine the required code points
 - ⦿ Document reference and reason for inclusion
 - ⦿ Review code points which are not analyzed

Example

Item #	Unicode Code Point	Glyph	Name and GC	Some languages using the character	Language, with EGIDS value	Reference
1	0621	ﺀ	ARABIC LETTER HAMZA;Lo	Arabic, Urdu, Punjabi, Sindhi	1 Arabic	[RFC 5564]
			...			
3	0623	ﺀ	ARABIC LETTER ALEF WITH HAMZA ABOVE;Lo	Arabic, Malay, Torwali	1 Arabic	[RFC 5564]
				
81	06AE	ﻙ⋮	ARABIC LETTER KAF WITH THREE DOTS BELOW;Lo	L'Alphabet National du Tchad (ANT)	1 ANT	[ANT]

Exercise

Item #	Unicode Code Point	Glyph	Name and GC	Some languages using the character	Language, with EGIDS value	Reference
1						
2						
3						
4						
5						

A world map where the continents are defined by a network of white dots and thin white lines. The dots represent nodes, and the lines represent connections between them. The background is a solid orange color. The word "Variants" is written in white, bold, sans-serif font on the left side of the map.

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?

Variant Relationships and Types

- ⊙ Variants are symmetric
 - ⊙ $A = B \Rightarrow B = A$
- ⊙ Variants are transitive
 - ⊙ $A = B$ and $B = C \Rightarrow A = C$
- ⊙ 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

0641	ف	فب	بف	بفب	1 (06A7)
06A2	فا	فبا	بفا	بفبا	5 (0641)
06A7	ف	فب	بف	بفب	5 (0642)
0642	ق	قب	بق	بقب	6 (06A7)

0641	06A2	a	Used interchangeably in Africa for languages using Western (African) orthography
0641	06A7	b	
0641	0642	b	
06A2	06A7	b	
06A2	0642	b	
06A7	0642	a	Used interchangeably in Africa for languages using Western (African) orthography

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

Examples

- ⦿ 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 XML Specification

LGR XML Structure

- ⊙ “lgr” element has three sub-elements:
 - “meta”: all meta-data associated with the LGR, such as its authorship, what it is used for, implementation notes and references.
 - “data”: the substantive code point data
 - “rules” (optional): information on contextual and whole-label evaluation rules, if any
 - with any specific “action” elements providing the disposition of labels and their variants

LGR XML Structure

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

Demo of online LGR Tool

Engage with ICANN



Thank You and Questions

Reach us at: IDNProgram@icann.org

Website: icann.org/idn



twitter.com/icann



[gplus.to/icann](https://plus.google.com/icann)



facebook.com/icannorg



weibo.com/ICANNorg



linkedin.com/company/icann



flickr.com/photos/icann



youtube.com/user/icannnews



slideshare.net/icannpresentations