



ICANN NO. 51 | 12-16 OCTOBER 2014
LOS ANGELES

#ICANN51

15 October 2014

IDN Program Update

Sarmad Hussain
IDN Program Senior Manager

#ICANN51



Agenda

- **IDN Program** - Sarmad Hussain
- **Maximal Starting Repertoire (MSR)** – Marc Blanchet
- **Transitioning from MSR to LGR**– Asmus Freytag
- **LGR Builder** - Kim Davies
- **Community Updates**
 - **Need for Latin GP** – Cary Karp
 - **Update on Arabic GP** – Meikal Mumin
- **Q/A**

15 October 2014

IDN Program Overview and Progress

Sarmad Hussain

IDN Program Senior Manager

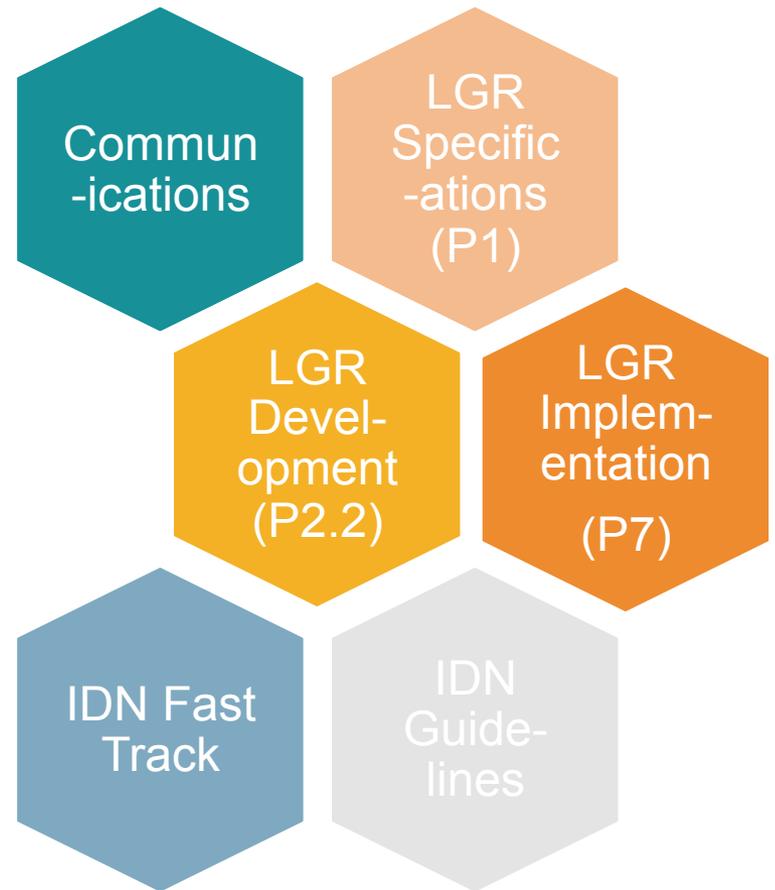
IDN Program Update

#ICANN51



IDN Program Overview

- Until recently Top Level Domains (TLDs) limited to a sub-set of ASCII in Latin script, e.g.: .com, .info, .br, .sg
- Now possible in different scripts – referred to as Internationalized Domain Name (IDN) TLDs, e.g.: .中国, .рф, .شبكة, .இலங்கை, .삼성
- IDN Program supports and undertakes various projects and programs related to IDN TLDs



IDN Program Overview

- **IDN TLD Program – Label Generation Ruleset (LGR)**
 - Determine validity and variants of an IDN label for the Root zone
- **IDN Fast Track Process Implementation**
 - Assist IDN ccTLD string evaluation for eventual delegation
- **IDN Implementation Guidelines**
 - Promote IDN registration policies and practices to minimize consumer risk and confusion
- **Communications and Outreach**
 - Update and engage community with the IDN Program

IDN TLD Program

#ICANN51



TLD Labels are Special - ASCII and IDN

- ASCII

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

-
- Traditional rules for Top Level Domain (TLD) labels
 - Letter Principle: ASCII [a-z]
 - Maximum length = 63 (in practice much shorter)

- IDN

- Rules for Internationalized Domain Name (IDN) Labels
 - **Valid U-Label:** Only PVALID and CONTEXT O/J code points and other constraints per IDNA 2008
 - **Valid A-Label:** Maximum A-label length = “xn--” + 59 chars for Punycode of U-label (RFC 3492)

- Rules for IDN TLD Labels
 - “Letter” principle for U-Label ???
 - Valid A-Label

Need Rules for Defining IDN TLD Labels

1. Which characters can form a label?
2. Which characters are confusable or variants?
3. What are additional constraints on labels?

IDN TLD Program

- To support IDNs and variants in the root zone,
 - the ICANN community, at the direction of the Board¹, 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>

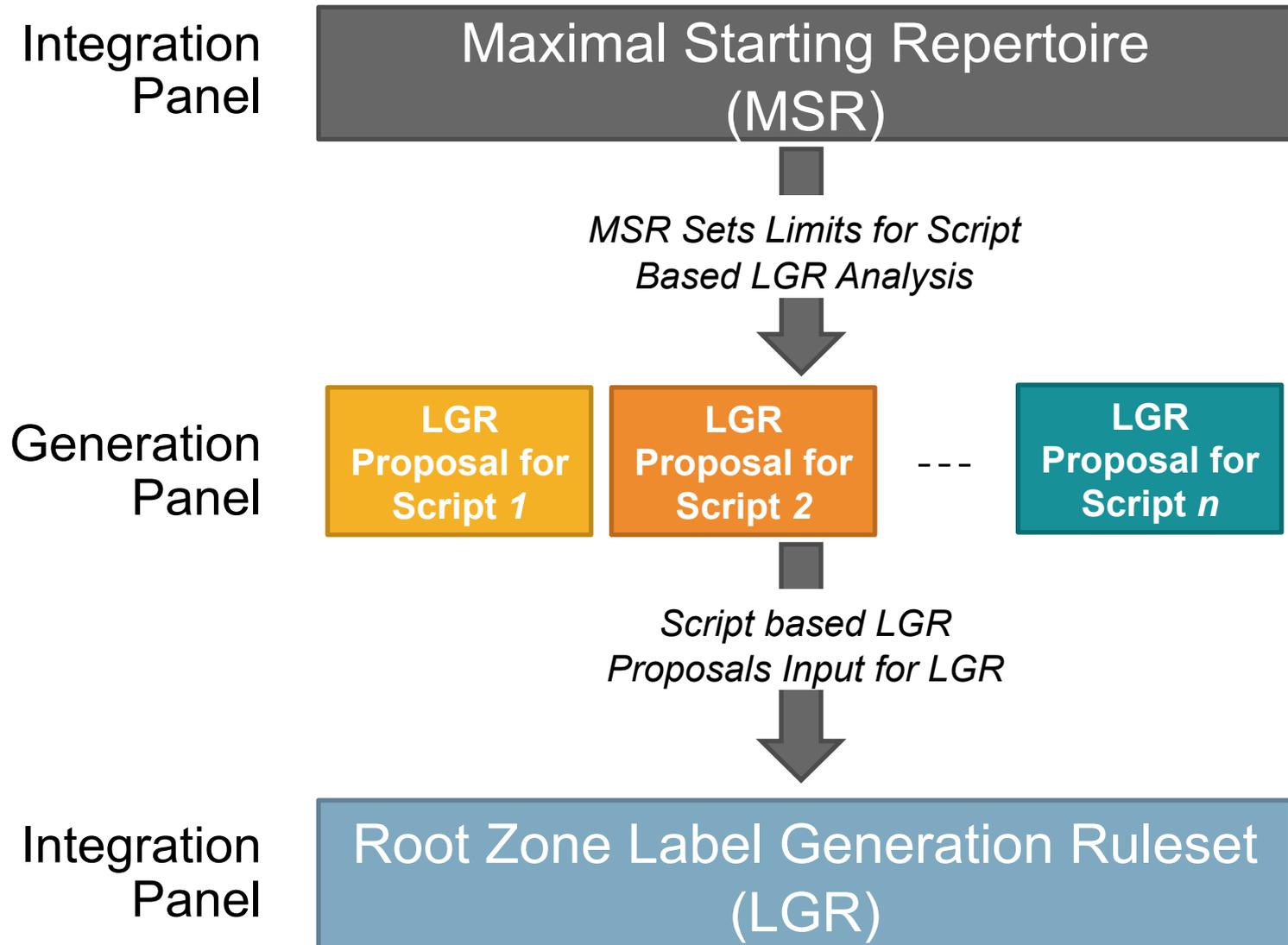
Work Completed under IDN TLD Program



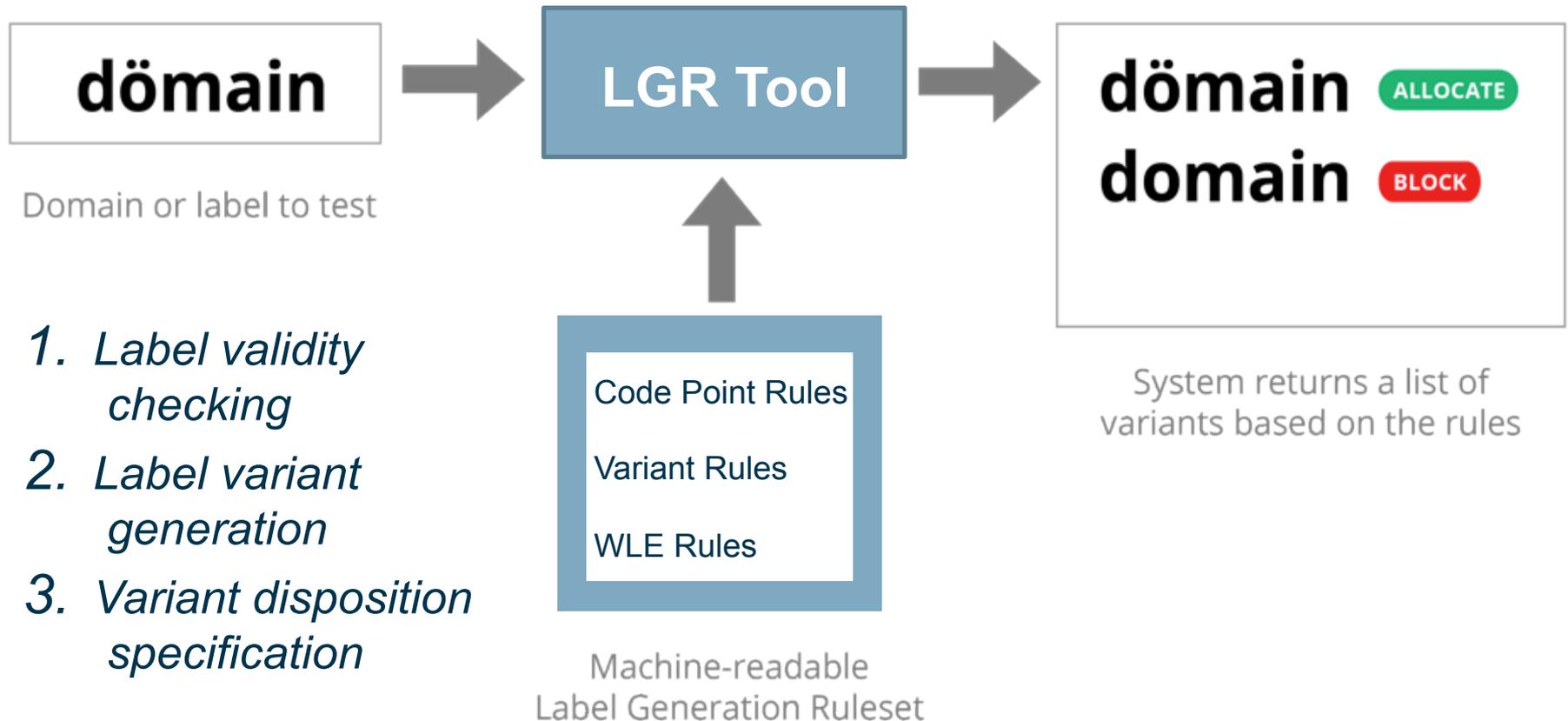
Community agreed to define a Label Generation Ruleset (LGR)

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

LGR Development Process



LGR Specification and Tool



<https://datatracker.ietf.org/doc/draft-davies-idntables>

LGR Development Progress

Jun
13

- Call for Integration Panel to develop Root Zone LGR

Jul
13

- Call for Generation Panels to propose Root Zone LGR

Sep
13

- Integration Panel formed

Feb
14

- Arabic Generation Panel Seated

Jun
14

- MSR-1 Released

Sep
14

- Chinese Generation Panel Seated

Dec
14

- MSR-2

Jan
14

- Arabic and Chinese Proposals for LGR

June
15

- LGR-1



Community Based GP Status

- Speak up for your Language (and Script)
 - 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
- Contribute by emailing interest to identlds@icann.org

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

Seated

Meeting

Forming

No Progress

Communications and Outreach

#ICANN51



Direct Community Engagement

- IDN Program Sessions at ICANN meetings
- IDN Program Updates to SOs/ACs at ICANN meetings
- APTLD Meeting, May, Muscat
- APriIGF, August, Delhi
- IGF, September, Istanbul
- TLDCON, September, Baku
- APTLD Meeting, September, Brisbane
- Email Communication to SOs/ACs – call to action

Community Outreach

- Blogs
 - <http://blog.apnic.net/2014/09/30/speak-up-for-your-language/>
- ICANN Community Wiki
 - <https://community.icann.org/display/croscomlgrprocedure/Root+Zone+LGR+Project>
- ICANN Website
 - www.icann.org/topics/idn
- ICANN Mailing Lists
 - {vip,lgr,ArabicGP, ChineseGP, ...}@icann.org
- Print and Social Media
- IDN Program Videos
 - <http://youtu.be/hPeKfiS7MNU>
 - <http://youtu.be/wnauGpYh96c>

15 October 2014

MSR Development

Presented by: Marc Blanchet

Integration Panel Member

IDN Program Update

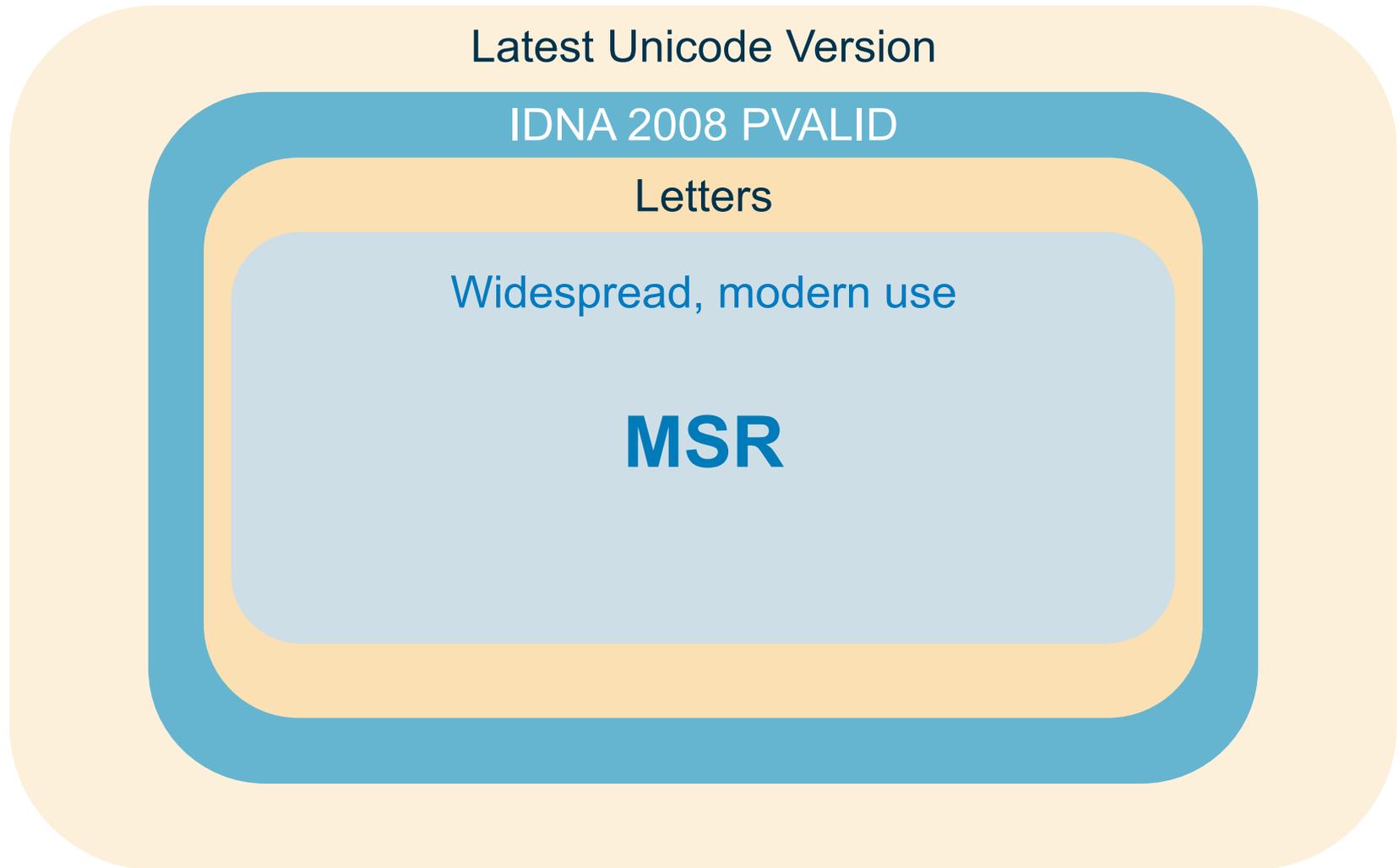
#ICANN51



MSR Process Summary

1. Start with latest [Unicode repertoire](#)
2. Limit to PVALID code points from latest [IDNA 2008 registry](#)
3. Apply “[Letter Principle](#)”
4. Limit to modern scripts targeted for the given MSR version
5. Exclude code points that are *unambiguously* limited to
 - Technical, phonetic, religious or liturgical use
 - Historical or obsolete writing systems
 - Writing systems not in widespread, common, everyday use ([EGIDS](#))
6. Result is a starting set from which the GPs pick their repertoire

Narrowing down the MSR Repertoire



Expanded Graded Intergenerational Disruption Scale

- EGIDS is an evaluation of language vitality
- Used as proxy for “effective demand” for the writing system
 - Not based on population size, but on “established vitality”
 - Not a perfect correlation with script use, but a useful criteria
 - Some writing systems are not stable or not widely used, even if the language isFor the MSR the IP uses the cut-off between Level 4 and Level 5
- *4: Educational*
 - Language in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education
- *5: Developing*
 - Language in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable

MSR-1: Foundation for initial round of GPs

- MSR-1 released on 20 June 2014:

<https://www.icann.org/news/announcement-2-2014-06-20-en>

- Work can now proceed for 22 scripts to create LGRs for the Root Zone
- Generation Panels will:
 - pick repertoire from within the MSR
 - decide whether code point variants exist
 - decide whether these should lead to allocatable or blocked variant labels
 - generate an LGR proposal for public comment and review (and integration) by Integration Panel

MSR-1 Content

Script	Count	Script	Count
Arabic	239	Hiragana	89
Bengali	64	Kannada	68
Cyrillic	93	Katakana	92
Devanagari	91	Lao	53
Georgian	37	Latin	305
Greek	36	Malayalam	73
Gujarati	66	Oriya	66
Gurmukhi	61	Sinhala	79
Han	19,850	Tamil	49
Hangul	11,172	Telugu	67
Hebrew	46	Thai	71
<i>Common</i>	0	<i>Inherited</i>	21
Total:		32,790	

- **Starting Point:**
Unicode 6.3 with
137,473 code points
- Of these 97,973 are **PVALID/CONTEXT**
per IDNA 2008
- Of these 32,790 are
in **MSR-1** as eligible
for the Root Zone.

MSR-2 status

- MSR-2 is cumulative, no deletions from [MSR-1](#)
- MSR-2 completes the repertoire by adding:
 - Armenian, Ethiopic, Khmer, Myanmar, Thaana, and Tibetan
- Content in numbers (as projected):
 - 28 scripts (up from 22 in MSR-1)
 - ~ 33,500 code points (~700 added to MSR-1)
- Schedule:
 - Release for public comment by end of 2014
 - Available in 2015

MSR-2 and IDNA 2008 registry

- MSR-2 is formally based on [Unicode 7.0](#)
- Content limited to [Unicode 6.3](#) subset until the IDNA 2008 registry is updated to Unicode 7.0 at IANA
 - <http://www.iana.org/assignments/idna-tables>
- This affects the potential repertoire for:
 - MSR-1 scripts (Arabic, Devanagari, and Telugu)
 - MSR-2 new script (Shan subset in Myanmar)
- Code points from 7.0 are considered candidates for addition if a new IANA table is published in time

15 October 2014

Transitioning from MSR to LGR

Asmus Freytag
Integration Panel Member
IDN Program Update

Developing an LGR – A Summary

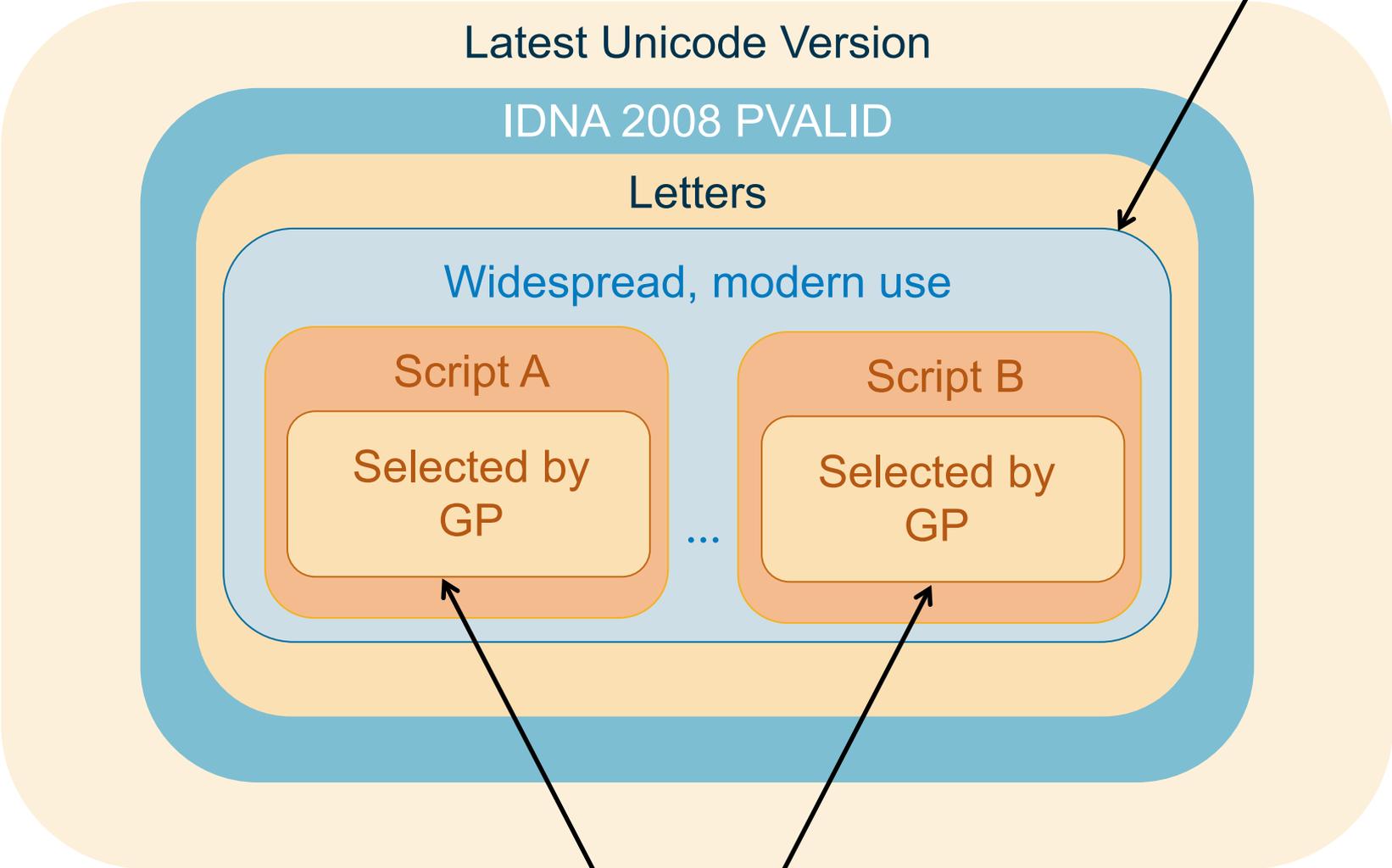
1. Start with the most recent MSR
2. Create a repertoire based on selected script
3. Are variants needed for the script?
 - a) define variant relations
 - b) decide which variants should lead to **blocked** vs. **allocatable** labels
4. Must some labels be prohibited via Whole Label Evaluation rules (WLE) ?
5. Document the decisions and their rationale
6. Create a XML file for repertoire, variants and WLE
7. Submit for public comment and IP review

Creating a Repertoire

1. Start with the latest [MSR](#)
2. Select a script (defined in scope of GP)
3. Select code points needed for IDN TLDs
 - Based on Principles laid out in the [Procedure](#)
 - Avoid any significant systemic risks
 - Avoid geographical or language bias
4. Document rationale for inclusion
 - Cover each code point or collection of code points
 - Show how selection satisfies the Principles

Building up the LGR Repertoire

MSR



LGR

Are Variants Needed?

- Does the repertoire contain variants?
 - Not all repertoires do. Some GPs will answer “no”.
- Variants are
 - Code points that are “the same” as other code points or code point sequences in the minds of users
- Two types of variants
 - Some lead to labels that are blocked, so two labels that are “the same” can’t be allocated separately
 - Some lead to labels that can both be allocated to the same applicant
- See the [Variant Rules](#) document for more detail

Defining Variants

1. Define which two code points are variants: **A ~ B**
2. Create the variant code point mappings
 - must be ***symmetric***: **A → B ⇒ B → A**
 - and ***transitive***: **A → B and B → C ⇒ A → C**
3. Assign a type to each mapping
 - Default types are “blocked” and “allocatable”
 - These work with predefined default actions
 - Other types are able to reduce number of allocatable variant labels
 - These require specific actions that finally evaluate to either “blocked” or “allocatable” variant labels

Constraints on Variants

- Procedure : “maximize number of blocked variants, and minimize the number of allocatable variants”
 - Steep complexity cost to allocatable variants
 - Limit to what is required for the writing system
 - Blocked variants can make LGR more robust

Are WLEs needed?

- Should some sequences of code points be generally prohibited within a script?
 - do they cause rendering problems? (in complex scripts)
- Would such a prohibition satisfy the Principles?
 - is increase in complexity offset by reduction in risk?
- Most Generation Panels would answer “no”
 - Advance discussion with the IP is recommended if a GP is considering WLEs

Defining WLEs

- WLEs are akin to regular expressions
- Labels matching a WLE are blocked
- Creating an WLE requires
 1. Defining one or more “rules”
 2. Optionally, defining “classes” of code points
 - Explicitly listed or based on Unicode properties
 3. Defining an “action” that causes the label to be “blocked” when a rule is matched
- The MSR contains a Default WLE which serves as an example

Indic Syllables

Rule 1: V[m]

Rule 2: {CnH}Cn[v][m]

Rule 3: CnH (at end)

where Cn is C[n]

Requirements:

- 1 : H must follow Cn
- 2 : m must not follow H
- 3 : v must follow Cn
- 4 : n must follow C

उ = V

0909

ऊ = Vm

0909, 0901

क = C

092B

क = Cn

092B, 093C

क = CnH

092B, 093C, 094D

कज = CnHC

092B, 093C, 094D, 091C

कजी = CnHCv

092B, 093C, 094D, 091C, 0940

Potential candidates for WLE

Documenting the LGR

- The GP documents the rationale for
 - Choice of repertoire, coverage and contents
 - Necessity, choice and type of variants
 - Necessity and design of WLEs
- The documentation will evaluate the design choices in light of the Principles
- See [Requirements for LGR](#) for more details

Formal Definition for LGR

- The formal definition of the LGR is an XML file
 - Format described in [Requirements for LGR](#)
- Plus: examples of labels, variant labels and labels blocked by WLEs
 - Only needed if the LGR contains variants or WLEs
- Optional: informative charts of the LGR repertoire
 - For example, like the annotated PDF files in the MSR

Submission for Review

- Submission format must match [Requirement for LGR](#)
- Completed LGR proposal is submitted to
 - ICANN for release for public comment
 - IP for review
- After comments are resolved
 - IP will attempt to integrate the LGR
 - If successful, the IP will submit an integrated LGR for public comment (consolidating several scripts)
 - If not, the IP will reject the proposal and tell the GP which features prevented integration

Throughout the Process

- Keep the Integration Panel in the loop
 - IP can only approve or reject the LGR proposal as a whole
 - Early discussions reduce the chance that some detail will lead to rejection
- Follow the Procedure
 - It is the authoritative prescription
 - The LGR Proposal must be compatible with its Principles

Throughout the Process (cont'd)

- Considerations for Designing an LGR for the Root
 - provides a helpful list of items for GPs to consider during the process
- Coordinate with GPs for Related Scripts
 - Where scripts are related, the IP will look for consistent and compatible treatment
 - If GPs coordinate among themselves, less chance that the IP will run into issues during Integration
 - Goal: consistent user experience

What should be Coordinated?

- Repertoire
 - Consistent treatment of similar repertoires
 - Examples: Indic scripts
- Variants
 - Compatible definition of variants
 - Examples: Han
 - Cross-script homoglyphs
 - Examples: Latin, Greek, Cyrillic
- WLE
 - Consistent treatment of structurally similar scripts
 - Examples: Indic scripts, definition of matra

Resources

- Considerations for Designing a Label Generation Ruleset for the Root Zone
 - <https://community.icann.org/download/attachments/43989034/Considerations-for-LGR-2014-09-23.pdf>
- Maximal Starting Repertoire (MSR-1)
 - <https://www.icann.org/news/announcement-2-2014-06-20-en>
 - <https://www.icann.org/en/system/files/files/msr-overview-06jun14-en.pdf>
- 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>
- Representing Label Generation Rules in XML
 - <https://tools.ietf.org/html/draft-davies-idntables>
- Requirements for LGR Proposals
 - <https://community.icann.org/download/attachments/43989034/Requirements%20for%20LGR%20Proposals.pdf>
- Variant Rules
 - <https://community.icann.org/download/attachments/43989034/Variant%20Rules.pdf>

15 October 2014

LGR Builder

A Tool for Generation Panels and the
Community

Kim Davies

Director Technical Services, IANA

IDN Program Update

Goals

- Interactive interface to build LGR files
 - Files are formatted using XML
 - XML is not always easy to author manually
 - Authoring tool can prevent/catch errors in compilation

Recap: What is an LGR?

- Machine-readable definition of registry policy regarding:
 - Allowable code points
 - Variant rules for specific codepoints
 - Whole-label rules
 - Processing rules (allocation, blocking, etc.)
- Designed so that with an LGR, a registry can automatically apply policies

XML files

Graphical User Interface

Web based application for intuitively constructing LGR tables or manipulating existing LGRs (“LGR Builder”)

Command Line Tools

Tools used to perform simple or bulk operations, such as verifying a set of strings against an LGR

Registry Systems

Domain registries implementing LGRs take supplied strings and check if they can be registered

IANA Systems

New table repository accepts LGRs, validates them on submission. Uses Root LGR to manage the set of TLDs in the root.

Software Libraries

Implements primitives for loading, saving, validating, modifying and using LGR tables

Python

Java

C

....

Phase 2 (Desirable functionality)

- Comprehensive UI that allows intuitive construction of LGRs
- Ability to test tables
- More complex features of LGR: tagging, WLE rules, etc.

Creating Russian table

а ✓	б ✓	в ✓	г ✓	д ✓	е ✓	ж ✓	з ✓	и ✓
й ✓	к ✓	л ✓	м ✓	н ✓	о ✓	п ✓	р ✓	с ✓
т ✓	у ✓	ф ✓	х ✓	ц ✓	ч ✓	ш ✓	щ ✓	ъ ✓
ы ✓	ь ✓	э ✓	ю ✓	я ✓	ё ✓	ђ ✓	ѓ ✓	є ✓
ѕ ✓	і ✓	ї ✓	ј ✓	љ ✓	њ ✓	ћ ✓	ќ ✓	ў ✓

Next steps

- Developing draft requirements
- Consultation with community to ensure it is useful
- Identify developers and execute
- Tool will be open source

15 October 2014

Developing a Latin Script IDN Repertoire for the Root Zone of the DNS

Cary Karp

Latin Generation Panel

IDN Program Update

Latin Script

- The Latin alphabet is used in the writing systems of more languages than is any other single script
- Only a small number of languages are written exclusively with its basic twenty-six letters
- The total requirement for additional forms is extensive

Two Base Letters of Many

o ò ó ô õ ö ø ò
õ ố ơ ố ơ ơ ớ ờ
ô õ ã ó ô õ ã ã ò
ố ơ ỏ ố ờ ỏ ỏ ộ
ớ ờ ỡ ỡ ợ ơ ơ ợ

h ĥ ħ ĥ
ħ ĥ ħ ħ
h h h h
h h h h
h y y y

Multiple Perspectives on Them

- Distinctions between these forms are of fundamental significance to the communities that use them

AND

- Communities that use the same characters often treat them in different ways

IDN Policies are Context Dependent

- Local considerations are relatively easy to accommodate in IDN policies for a zone that has a clear language nexus
- Even then, it may be necessary to limit the orthographic nuance that can be applied

ASCII Still Goes a Long Way

- If forgoing a diacritical mark does not change meaning it is simple enough to accept the undecorated base letter

naivete

naïveté

But Doesn't Even Fully Cover English

- The situation is not quite the same when a diacritical mark is contrastive

resume

résumé

Much Less Swedish

- Yet another situation applies when the alternative might be the substitution of two base characters for a single marked one

nörden
noerden
norden

Or Personal Names

- This can be critical in the representation of proper names

goethe
göthe

Special Conditions Apply to the Root Zone

- The root zone has no intrinsic language attributes and serves all language communities
- Its IDN policies need to accommodate all communities in as equitable a manner as possible

Getting from Here to There

- The final report of the VIP Study Group for the Latin Script root includes a list of the entire repertoire of Latin code points that can be used in an IDN label
- The LGR Integration Panel reduced that list to a Maximum Starting Repertoire by excluding code points with properties that are not appropriate for the root zone
- Neither document is intended to meet the IDNA protocol requirement for the operator of a zone — in this case the root — to collate a subset of the available repertoire appropriate for its target community
- A Script Generation Panel is needed to address remaining concerns and propose a Latin code point repertoire specifically for the root zone

15 October 2014

Update on Arabic GP

Meikal Mumin

Task Force on Arabic Script IDNs / Arabic GP

IDN Program Update

Community driven way forward: **Task Force on Arabic Script IDNs**

- Creation and oversight by community based Middle East Strategy Working Group (MESWG;

<https://community.icann.org/display/MES/MESWG+Members>)

- TF-AIDN Objectives: a holistic approach
 - Arabic Script Label Generation Ruleset (LGR) for the Root Zone
 - Second level LGRs for the Arabic script
 - Arabic script Internationalized Registration Data
 - Universal acceptability of Arabic script IDNs
 - Technical challenges around registration of Arabic script IDNs
 - Operational software for registry and registrar operations
 - DNS security matters specifically related to Arabic script IDNs
 - Technical training material around Arabic script IDNs

Membership

- Currently **29 members** – applications still being received
- From **17 countries** – Australia, Bahrain, Egypt, Ethiopia, Germany, Iran, Jordan, Lebanon, Malaysia, Morocco, Pakistan, Palestine, Saudi Arabia, Sudan, U.A.E., U.K., U.S.A.
- Members of **nine language communities using Arabic script** – Arabic, Malay, Saraiki, Sindhi, Pashto, Persian, Punjabi, Torwali, Urdu, with further **expertise** in use of Arabic script from East Asia, South Asia, Middle East, and Africa
- Coming from **diverse disciplines** – academia (linguistics and technical), registries, registrars, national and regional policy bodies, community based organizations, technical community

How to reach out to the Task Force on Arabic Script IDNs?

- Membership open, community based
- Details and interests of members posted by MESWG
- Discussions publicly archived
- Details at <http://lists.meswg.org/mailman/listinfo/tf-aidn>
- Background and introduction to TF-AIDN
 - <https://community.icann.org/display/MES/Task+Force+on+Arabic+Script+IDNs>
- Workspace, news and document archive
 - <https://community.icann.org/display/MES/TF-AIDN+Work+Space>
- Email archive
 - <http://lists.meswg.org/pipermail/tf-aidn/>

Arabic Script TLDs Assigned or Delegated

1. الجزائر

2. عمان

3. ايران

4. امارات

5. بازار

6. باكستان

7. الاردن

8. بهارت

9. المغرب

10. السعودية

11. السودان

12. مليسيا

13. شبكة

14. سورية

15. تونس

16. مصر

17. قطر

18. فلسطين

19. عراق

20. بازار

21. موقع

Progress - Outreach to the community

- Launch at the Arab IGF Meeting in Algiers
- Presentation during the IGF in Bali
- Outreach during the ME DNS Forum
- Presentation to the community at ICANN Singapore
- Presentation to the community at the APTLD Meeting
- Presentation to the community at ICANN London
- Presentation to the community at IGF Istanbul
- Submitted a proposal for a workshop at Arabic IGF Lebanon

Progress –

Work accomplished / challenges encountered

- Work accomplished
 - Formation of Arabic Script Generation Panel
 - Establishing principles for inclusion, exclusion, and deferral of Unicode code points
 - Analysis of Maximum Starting Repertoire (MSR) and feedback to Integration Panel (IP)
 - Analysis of code points for Label Generation Rules (LGR)
- Issues encountered mainly due to
 - Representation of scripts as a proxy of language
 - Lack of data on orthographies
 - Lack of representation of script/language communities

Progress - Summary of code points

Description	Number	Language / Example	Glyph
Total number of code points considered	343		
Code points PVALID according to IDNA 2008	289		
Code points included in MSR-1	239		
Code points for which TF-AIDN found evidence of use	167		
Specific code points shortlisted but excluded in IDNA 2008	1	Jawi / U +0662 –	ﻯ
Specific code points shortlisted but excluded in MSR-1	5	Sindhi / U +06FD	ﻮ
Code points where TF-AIDN found evidence of use but which are out of scope based on general MSR rules (Threatened or declining language, optional character)	32	Ormuri / U +076B Kurdish / U +0692	ﻮ
Number of code points suggested for LGR by TF-AIDN	129		

Progress - Next steps

- XML representation
- Finalize the discussions on variants
- Whole Label Rules – Aug – Oct 14
 - Document principles for whole label variants
 - Define whole label variants
 - Release for public comments
- Finalization – Nov –Dec 14
 - Finalize LGR for Arabic script
 - Submit to ICANN/IP
 - Release for public comments

ٲرٲما كاسٲه

شكرآ

شكرٲه

باسٲاس

تشكر

Thank you

Questions & Answers

USEFUL LINKS:

- LGR Procedure:
<http://www.icann.org/en/resources/idn/variant-tlds/lgr-procedure-20mar13-en.pdf>
- MSR-1: <https://www.icann.org/news/announcement-2-2014-06-20-en>
- Call for Generation Panels:
<http://www.icann.org/en/news/announcements/announcement-11jul13-en.htm>
- Setting up and running a Generation Panel:
<https://community.icann.org/display/croscomlgrprocedure/Generation+Panels>
- Community Wiki for LGR Project :
<https://community.icann.org/display/croscomlgrprocedure/Root+Zone+LGR+Project>
- IDN Variants:
<https://www.icann.org/resources/pages/variant-tlds-2012-05-08-en>
- To submit expressions of interest, or if you have additional questions, please contact ICANN at: identlds@icann.org

Engage with ICANN on Web & Social Media



twitter.com/icann



gplus.to/icann



facebook.com/icannorg



weibo.com/icannorg



linkedin.com/company/icann



flickr.com/photos/icann



youtube.com/user/ICANNnews



icann.org