ICANN

NO. 49
23-27 MARCH 2014

Singapore

IDN Root Zone LGR Workshop: Integration & Generation Panels

Wednesday, 26 March 2014

#ICANN49

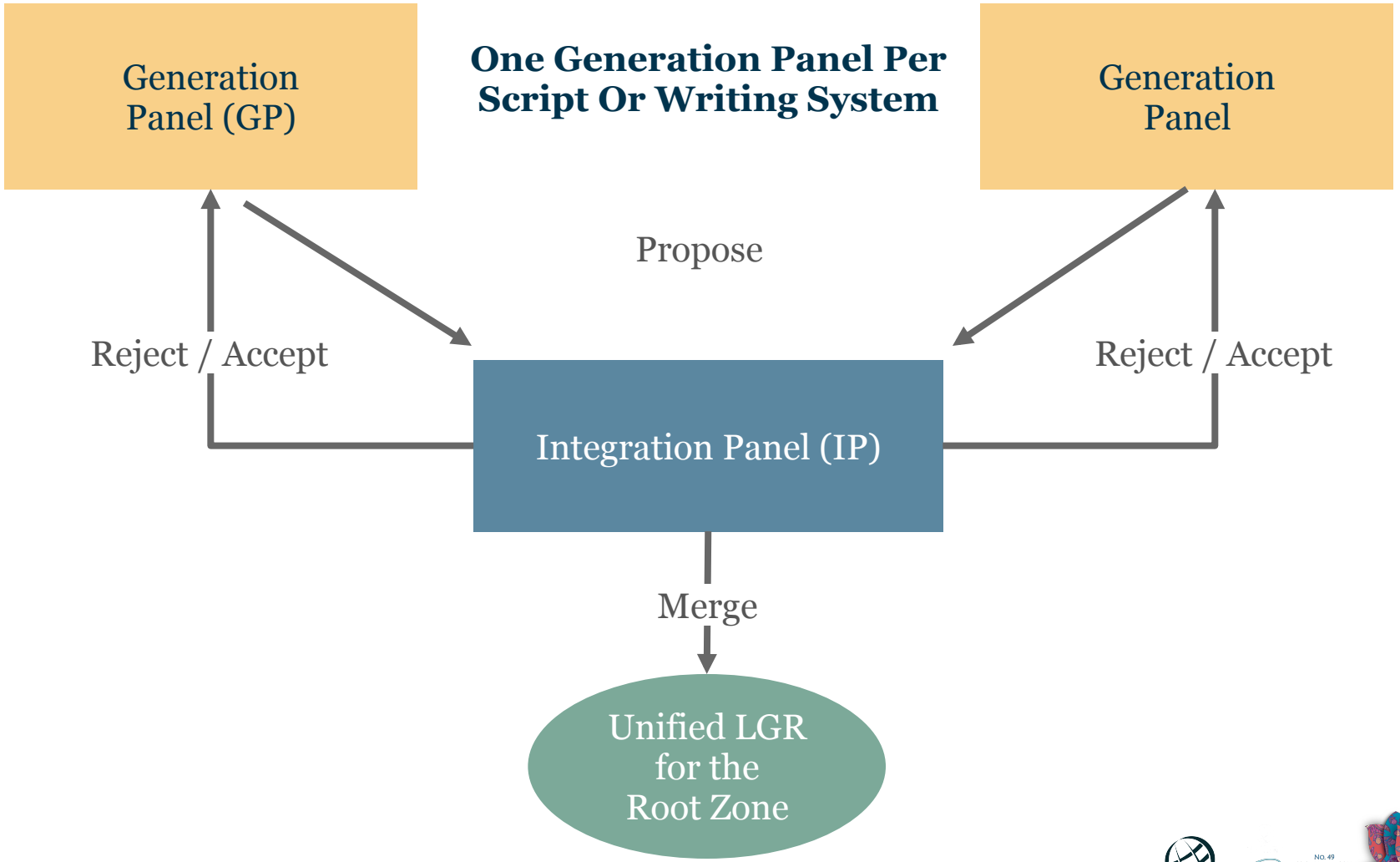
Agenda

Topic		Allocated Time
Welcome and Introductions		13.00-13.15
Session 1	**Community's role in creating the LGR** **I. Guidance on how to form a Generation Panel and collaborate with the Integration Panel (15 minutes)** **II. Community work to establish GPs (90 minutes)** **a. Arabic Generation Panel** **b. Being formed: C, J and K** **c. Being formed: Neo-Brahmi**	13.15-15.00
Break		15.00-15.15
Session 2	**Maximal Starting Repertoire and** **Whole Label Evaluation Rules**	15.15-15.50
Session 3	**Training on XML format to represent the LGR**	15.50-16.50
Wrap up and close workshop		16.50-17.00

Session 1: Community's role in creating the Label Generation Rules (LGR)

Guidance on how to form a Generation Panel and collaborate with the Integration Panel

Developing the Root Zone LGR



Forming a Generation Panel

- ICANN issued call for Generation Panels
- Focal point for a given script / script family
  - Community-based
  - Community-driven
  - Community-established
- Members represent languages/regions
  - Expertise in Linguistics, Policy, Unicode, IDNA, DNS, Registry/Registrar, etc.
- Community secures funding for meetings, etc.
  - Submits GP proposal to ICANN to form Panel

Time To Get Involved

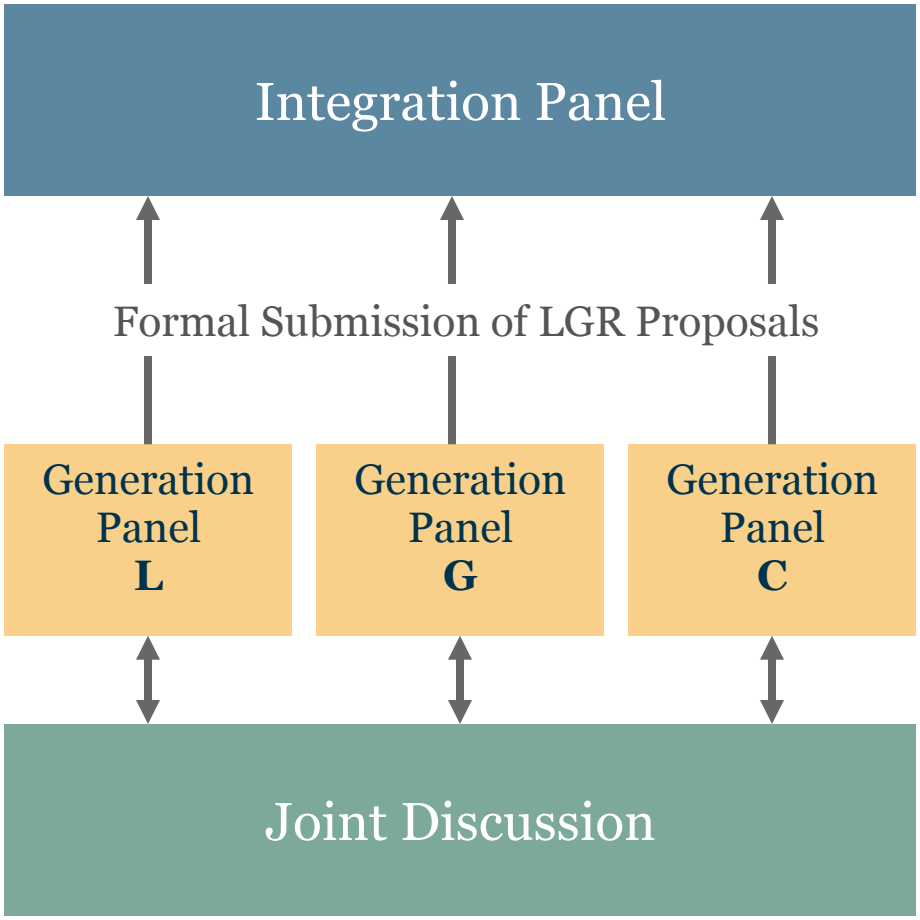
- Go to Project workspace and get documents
  - Form a generation panel
  - Volunteer to join a generation panel
- Contact ICANN staff for information
- Take part in public review of MSR
- Take part in public review of LGR proposals
- Take part in public review of integrated LGR
- Disseminate message to interested communities and community members

Which Panels Are Needed?

- MSR-1 covers 18 applied for scripts*
  - Additional 4 scripts are related and need to be considered simultaneously
  - For full coverage, need GPs for each of those 22 scripts
- Cyrillic, Greek, Georgian, Hebrew, Lao, Latin, Thai
  - No community input so far
- Arabic, Chinese, Indic, Korean, Japanese
  - From exploratory work to fully formed GP

* Japanese and Indic encompass multiple scripts

Related Scripts Example



- Work on a script cannot be finalized until all communities with closely related scripts have submitted proposals
- Generation Panels for related scripts should discuss and resolve joint issues before any of them submit their LGR proposals
- The joint discussion may be informal or formalized in a coordination committee, or one Generation Panel may cover the set of related scripts

Interactions Between IP and GP

- Codified in Section B.4.3 of the LGR Procedure
- MSR review & comments by GP (or proto GP)
  - MSR as fixed starting point after review
- GPs and IP to engage early to manage expectations
  - continuing dialogue and status updates
  - iron out potential points of disagreements early
- GPs with related scripts must coordinate
- Avoid surprises in LGR proposal
  - IP decisions must be unanimous
  - IP cannot reject part or modify a proposal (all-or-nothing)

Session 1: Community's role in creating the LGR

Community work to establish GPs

Arabic Generation Panel

Task Force on Arabic Script IDNs: Overview and Progress

ICANN Singapore Meeting

Task Force on Arabic Script IDNs (TF-AIDN)
Middle East Strategy Working Group (MESWG)
tf-aidn@meswg.org

History of Community Work on Arabic Script IDNs

- Very active Arabic script community; time line of Arabic script community
  - 2003: Arabic Domain Names Task Force (ADN-TF) by UNESCWA and LAS
  - 2005:Persian (Arabic script) SLD domain names offered
  - 2005: Arabic Domain Names Pilot Project
  - 2008: Arabic Script IDN Working Group (ASIWG)
  - 2009: ICANN IDN ccTLDs Fast Track Program initiated
  - 2010: Arabic Language in Internet Domains (RFC 5564)
  - 2010: First IDN ccTLDs delegated
  - 2011: Arabic Script Case Study for IDN Variants in the Root
  - 2013: Task Force on Arabic Script IDNs by MESWG
- Significant expertise in a variety of relevant domains: Unicode, DNS, IDNA and Registry and Registrar Operations
  - Experience from SLD deployments in ccTLDs
  - Experience from IDN ccTLD deployments

IDN TLDs Assigned or Delegated

1. الجزائر
2. عمان
3. ايران
4. امارات
5. بازار
6. پاكستان
7. الاردن
8. بهارت
9. المغرب
10. السعودية
11. سودان
12. مليسيا
13. شبكة
14. سورية
15. تونس
16. مصر
17. قطر
18. فلسطين

Code Points for Arabic Script IDNs

- Arabic script has the following specifications
  - ISO 15924 code: **Arab**
  - ISO 15924 no.: 160
  - English Name: Arabic
- Relevant sub-set
  - The complete set of code points in the Arabic script fall in the following Unicode ranges
    - Arabic U+0600 – U+06FF
    - Arabic Supplement U+0750 – U+075F
    - Arabic Extended A U+08A0 – U+08FF
  - Additional code points to be considered (as per Arabic Variant Issues Report)
    - Zero Width Non-Joiner U+200C
    - Zero Width Joiner U+200D

Typology of Arabic IDN Variants

- Same
  - Identical — U+06CC (ی)/U+0649 (ى)
  - In Context — U+06A9 (کبک)/U+0643 (كبك)
  - Normalization — U+0632 (ز)/U+0631+U+06EC (ر + ◌ = ز)
- Similar
  - Character — U+06AA (ے)/U+06A9 (ک)
  - Diacritic — U+062A (ت)/U+067A (ٺ)
- Different
  - Shape — U+0629 (ة)/U+06C3 (ۃ)
  - Character — U+0629 (ة)/U+06C1 (ہ)

IDN Variants Needs and Challenges

Security and Stability Needs

پاکستان پاکستان

U+0643 U+06A9

xn--mgbai9a5eva00b xn--mgbai9azgqp6j

- 120+ cases of visually same or similar Arabic script characters identified by case study team
  - Variants must not be allocated independently
  - Variants may need activation to allow user access (w/ different KB)
- 16 IDN ccTLD applications with 4 applications with variants

Security and Stability Challenges

- Consistency and innumerability
  - Consistent across and within TLDs
  - Minimal activation for manageability
- Management tools
  - Registration
  - Configuration and Maintenance
  - Security and Monitoring
- Usability in applications
  - Browsing, emailing, etc.
  - Searching, privacy, etc.

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

Timeline of Formation

Aug 13
- Call for membership of TF-AIDN (released at Arab IGF II) by MESWG

Oct 13
- Formation of TF-AIDN by MESWG
- Initiation of Proposal Development for Formation of Generation Panel

Dec 13
- Informal consultation on final draft of proposal with ICANN/IP

Jan 13
- Submission of proposal to ICANN for formation of Arabic Script Generation Panel (GP)

Feb 13
- Acceptance of Proposal by ICANN → formation of Arabic Script GP
  - All members of TF-AIDN members of Arabic Script GP

Membership

- Currently **21 members** – applications still being received
- From **11 countries** – Egypt, Germany, Iran, Jordan, Lebanon, Malaysia, Morocco, Pakistan, Palestine, Saudi Arabia, Sudan
- Speaking more than **nine languages** – Arabic, Malay, Saraiki, Sindhi, Pashto, Persian, Punjabi, Torwali, Urdu, and African Languages
- With **expertise** in use of Arabic script from East Asia, South Asia, Middle East, North Africa, Africa
- Coming from **diverse disciplines** – academia (linguistics and technical), registries, registrars, national and regional policy bodies, community based organizations, technical community

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

Method of Work

- Open call for each work item within TF-AIDN
- Volunteers develop the work item
- Work item presented to TF-AIDN for discussion
- Work item finalized with consensus
- Discussion within TF-AIDN archived at public list
- All teleconferences recorded and posted at public wiki page of TF-AIDN under MESWG wiki page
- All materials finalized after a public comment period
- URLs
  - 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/

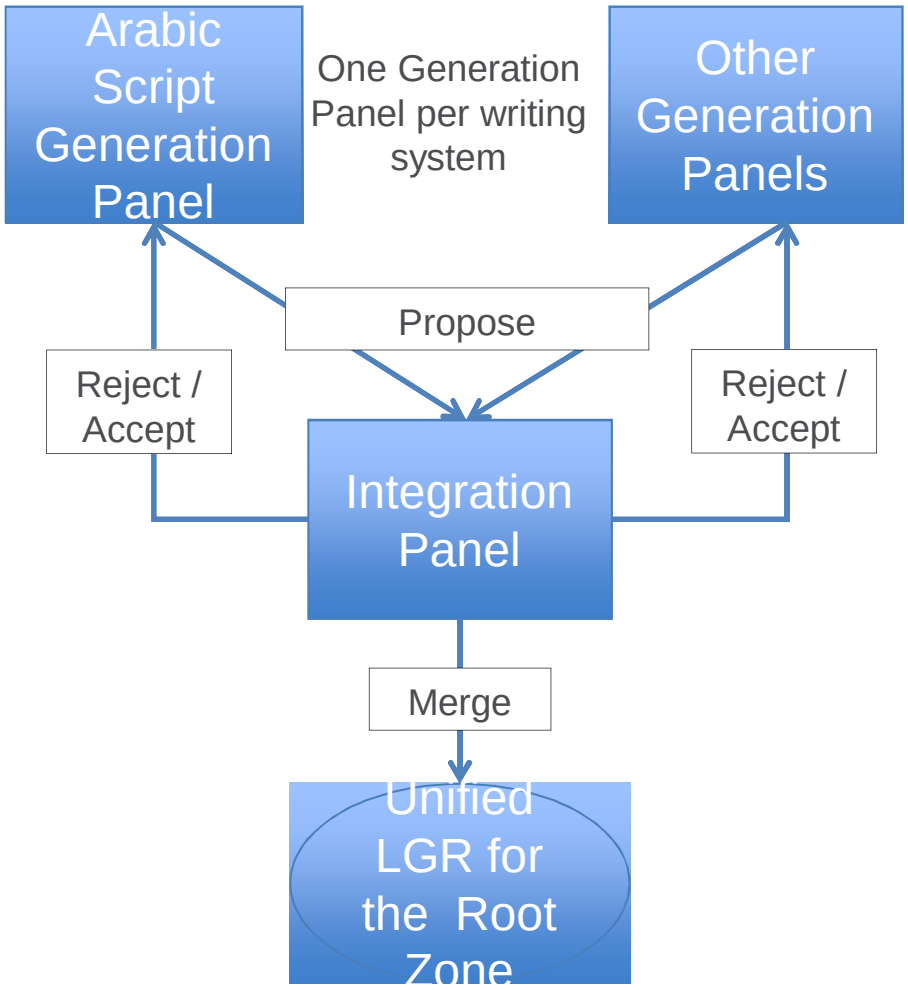
Proposal for Arabic Script GP

- Development of Proposal as per the ICANN guidelines
  - General Information
    - Target Script
    - Principle Languages
    - Countries with Significant Use
  - Composition of Panel
    - Panel Chair and Members
    - Panel Diversity
    - Relationship with Past Work
  - Work Plan
    - Suggested Timelines with Significant Milestones
    - Schedule of Meetings and Teleconferences
    - Sources of Funding
- Submission of proposal to ICANN

Role of GP in IDN Project

- Generation Panels generate proposals for script specific LGRs, based on community expertise and requirements

- Integration Panel reviews proposals and integrates them into common Root Zone LGR while minimizing the risk to root zone as shared resource



Current Work in Arabic Script GP

- Completed principles for inclusion, exclusion or deferral of code points for Arabic script LGR for the root zone - released for public comment
- Reviewing Code Points for Inclusion and exclusion from LGR
- Reviewing of MSR
- Discussing IDN Variants and defining general principles

Principles for Code Points

- Inclusion principles, e.g.:
  - *Letter code point* which is a *letter* and has *established contemporary use* in a language
  - *Mark code point* which represents a *required mark*, where at least one of the *letter* it forms has *established contemporary use* in a language
- Exclusion principles, e.g.:
  - Code point either deprecated or not recommended for use in Unicode Standard; exception being it meets one of the inclusion criteria with no alternative code point(s)
  - Code point specifically for historic use with no *established contemporary use*
- Deferral principles, e.g.:
  - Code point which can neither be confirmed for inclusion nor for exclusion based on principles
  - Such code points will be considered in future versions of LGR, when more concrete information is available

F2F Meeting in Singapore

- 10 participants funded by ICANN
- F2F Meeting Program: 20-22 Mar
  - Character repertoire (based on principles) [1.5 Days]
  - Review of general principles based [0.25 Days]
  - Review of MSR released by Integration Panel [0.5 Days]
  - Review of Variant and Whole Label Issues [0.75 Days]
- ICANN Meeting Program: 23-26 Mar
  - Training on IDN P1
  - Public sessions by TF-AIDN during the ICANN meeting
  - ICANN IDN Variant Program Update Session
  - Meetings with MESWG, Integration Panel, ICANN IDN Team, ICANN staff and policy makers

Next Steps for GP

- Characters – Jan-Apr 14
  - Finalize code point inclusion/exclusion principles
  - Determine code point for inclusion/exclusion/deferral
  - Release for Public Comments
- Variants – Apr – Jul 14
  - Document principles for variants
  - Define variants
  - Release for Public Comments
- 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

Middle East DNS Forum

- Feb. 2013 – Stakeholders from the region
- Presented issues pertaining to IDNs to the community
- Engaged stakeholders and introduce TF-AIDN and the work been undertaken
- Received community feedback; Priority
  - Arabic Script IDN Universal Acceptance
  - Arabic Script Email
- Enlisted candidates to join the TF-AIDN from the community

Next Steps for TF-AIDN

- Complete work on Arabic Script LGR
- Initiate additional work identified
  - Arabic Script Email
  - Universal Acceptability of Arabic Script IDNs

ICANN's Support for the TF-AIDN

- Facilitating the work of the TF-AIDN
  - Dedicated wiki space at https://community.icann.org/display/MES/TF-AIDN+Work+Space
  - Arranging for conference calls when needed, and posting call recordings on the wiki space
  - Provided support to the first F2F meeting in Singapore; both logistically and financially
  - Arranging for meetings with relevant ICANN staff when needed
  - Addressing issues and concerns raised by TF-AIDN members
- Financial support for the next TF-AIDN meeting scheduled for September 2014
- Ensuring public visibility of the TF-AIDN at relevant fora, gatherings, and events.

تریما کاسیه

شکراً

شکریہ

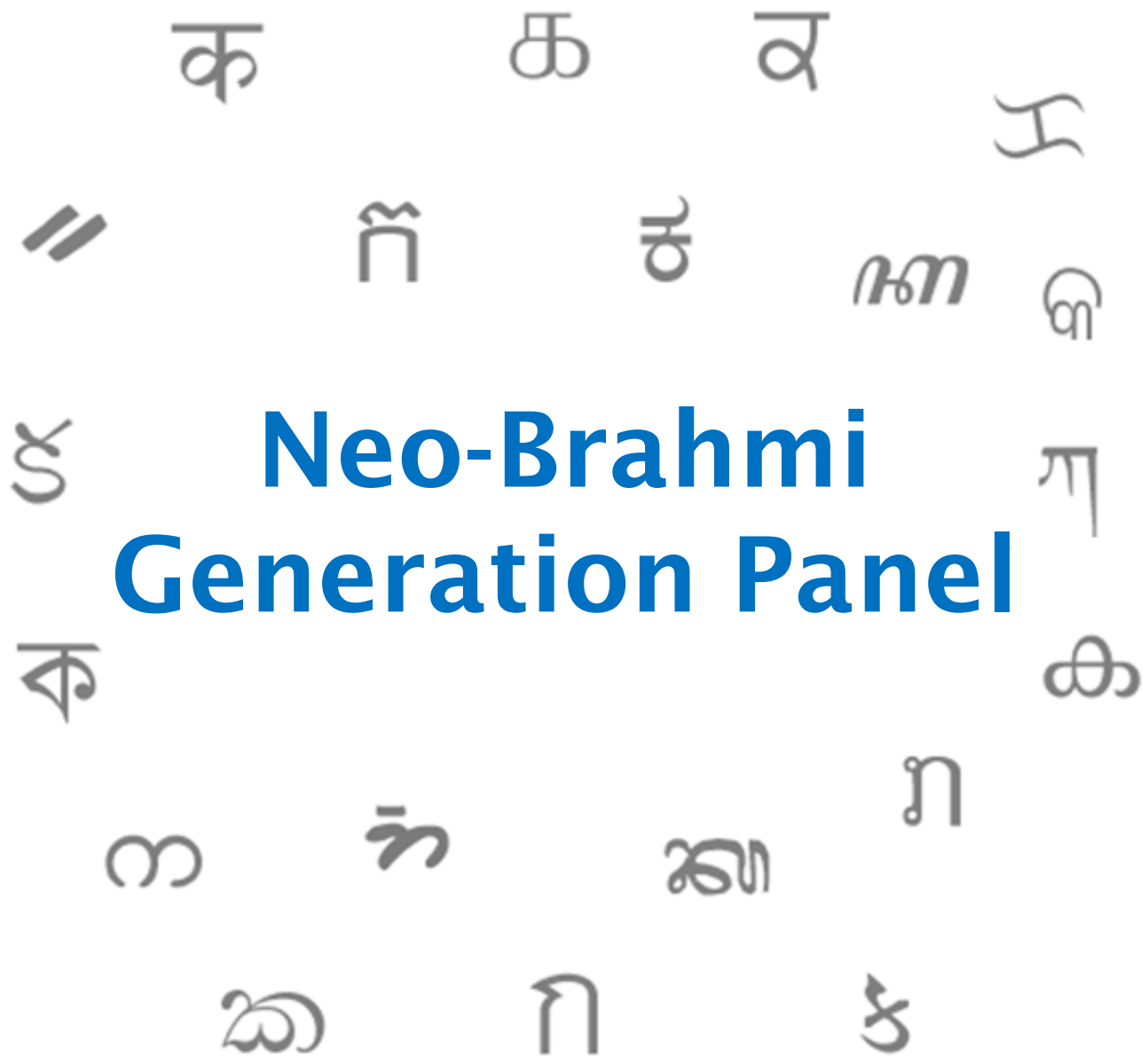
با سپاس

تشکر

Thank You

Being Formed:
C, J, and K Panels

Being formed:
Neo-Brahmi Panel



Neo-Brahmi Generation Panel

Overview

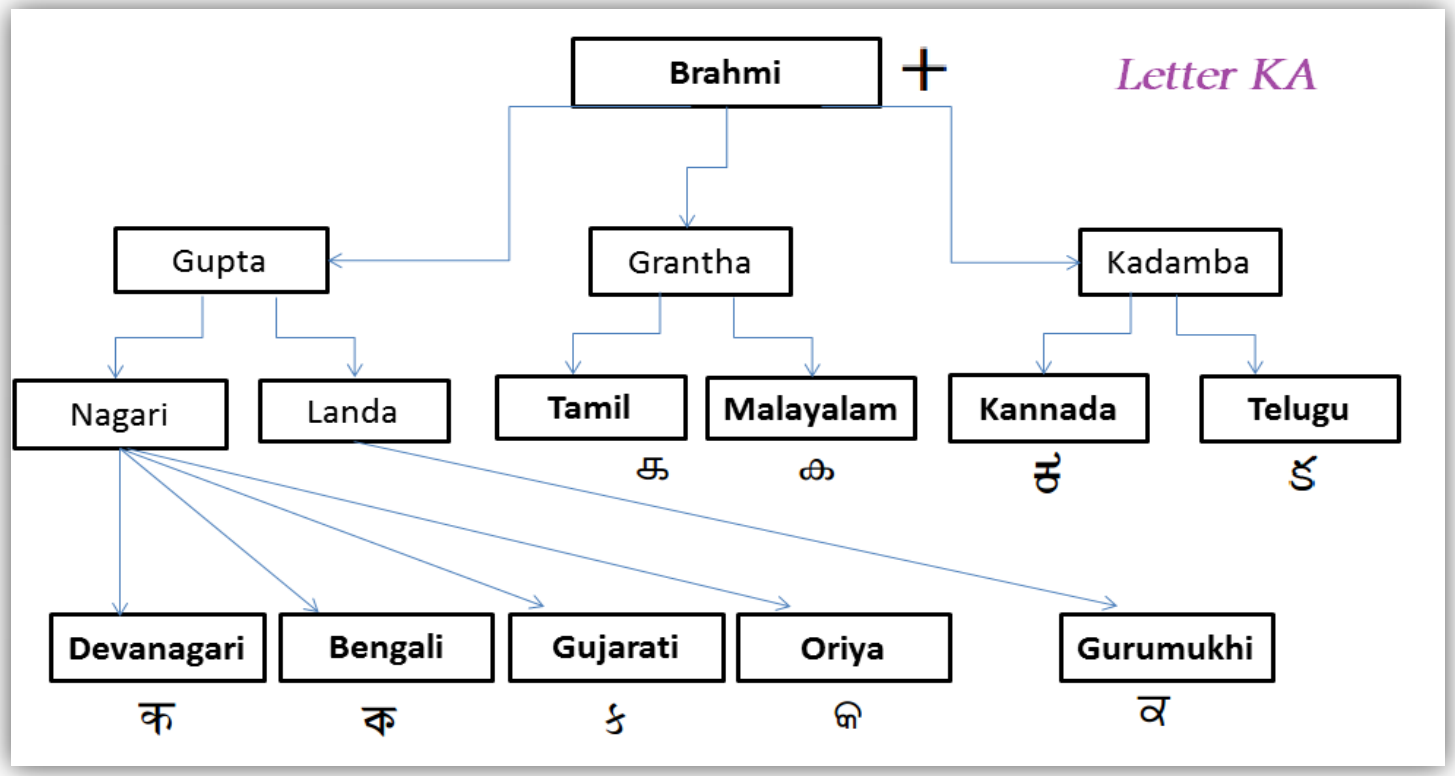
- What is Brahmi ?
- Why Brahmi ?
- Why Neo-brahmi approach?
- Brahmi derived scripts: Linguistic scenario
- On-going efforts & challenges

What is Brahmi?

- An ancient script which evolved during the final centuries BC.
- Most of the modern scripts in Indian subcontinent have been derived from Brahmi
- Geographically the scripts being used in Central Asia, South Asia and South-East Asia
- These scripts are used by multiple language families: Largely by Indo-Aryan and Dravidian in India

Why Brahmi?

- All the scripts derived from Brahmi are Abugida, also known as alphasyllabary.
- The lineage, example of “Letter Ka”



Why Brahmi?

- Despite their variations in the visual forms, the basic philosophy in their usage is common
- They all are “akshar” driven, and follow a specific syntax
  - Analogical reference can be made to Indian National standard, IS 13194:1991 – Section 8
- This syntax being the implicit foundation in representation of these scripts in the digital medium, adherence to the structure acts as a obligatory security consideration even in the case of Internationalized Domain Names.
- Here we are dealing with “root”

Why the term Neo-Brahmi?

- The term Neo-Brahmi in its linguistic sense refers to languages using a modified form of the matricial Brahmi script.
- Of all the scripts derived from “Brahmi”, not all are in modern usage
- Approach is in consonance with the conservatism principle of the LGR procedure

Linguistic Scenario:

Indian subcontinent

Language	Language Family	Script
Assamese	Indo-Aryan	Bangla (Modified)
Bangla	Indo-Aryan	Bengali
Bodo	Tibeto-Burman	Devanagari
Dogri	Indo-Aryan	Devanagari
Gujarati	Indo-Aryan	Gujarati
Hindi	Indo-Aryan	Devanagari
Kannada	Dravidian	Kannada
Kashmiri	Indo-Aryan	Perso-Arabic, Devanagari
Konkani	Indo-Aryan	Devanagari, Roman
Maithili	Indo-Aryan	Devanagari
Malayalam	Dravidian	Malayalam
Manipuri	Tibeto-Burman	Bangla, Meetei-Mayek
Marathi	Indo-Aryan	Devanagari
Nepali	Indo-Aryan	Devanagari
Odia (Oriya)	Indo-Aryan	Odia (Oriya)
Punjabi	Indo-Aryan	Gurmukhi, Shahmukhi
Sanskrit	Indo-Aryan	Devanagari
Santali	Munda	Devanagari, Ol Chiki
Sindhi	Indo-Aryan	Perso-Arabic, Devanagari,
Tamil	Dravidian	Tamil
Telugu	Dravidian	Telugu

Other Brahmi-Derived Scripts

- Cultures outside India have adopted the Brahmi script family to represent their languages.
- These belong to what Unicode classifies as Central Asian, South Asian and Southeast Asian scripts.
  - Link: http://www.unicode.org/charts/
- Some of them are Sinhala, Tibetan, Dzongkha, Myanmar and Thai

On-going efforts

- Identification of Experts in different areas (Community Representative, Linguistics, Registry/Registrar Operations)
- Familiarizing them with the LGR Procedure

Global Participation required

- In addition to the languages listed above there is a need of participation from other cultures sharing the languages derived from Brahmi.
- The list presented here is of countries where these languages have or share the status of “Official Language”
  - **Tamil** – Singapore and Sri Lanka
  - **Nepali** – Nepal
  - **Bengali** – Bangladesh
  - **Hindi** - Fiji

Challenges

- Generating interest in the community for working as volunteers
- Need for easy-to-explain form of documentation for the procedure and its expected outcome
  - Since sometimes “what to do” is not understood, “why to do” remains a question.
- At times lack of funding


THANK YOU
धन्यवाद

BREAK

15 min

Session 2:

Maximal Starting Repertoire & Whole Label Evaluation Rules

Context

- MSR (Maximal Starting Repertoire) and WLE (Whole Label Evaluation) are part of the Root Zone LGR (Label Generation Rules) development
  - Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels
- The LGR project will result in a set of rules:
  - Define permissible labels for the root
  - Can be mechanically applied (automated)
  - Expressed as a set of parallel and consistent per-script rules
  - The Integrated LGR applies to all scripts covered

Maximal Starting Repertoire

- One of the initial tasks of the Integration Panel
- Outer limit of code points allowed in the root zone
- Subset of IDNA 2008 PVALID Code points
  - No digits/punctuation
  - No context dependent / unstable
  - No historic / obsolete
  - No limited use

MSR Is Outer Limit

- Code points not in the MSR
  - Cannot be part of the Root Zone LGR
- MSR includes code points that may not be part of the final LGR
  - Some “require further evaluation” by GP
  - Some may ultimately not be acceptable
  - Some may be only allowed in specific sequences or conditions

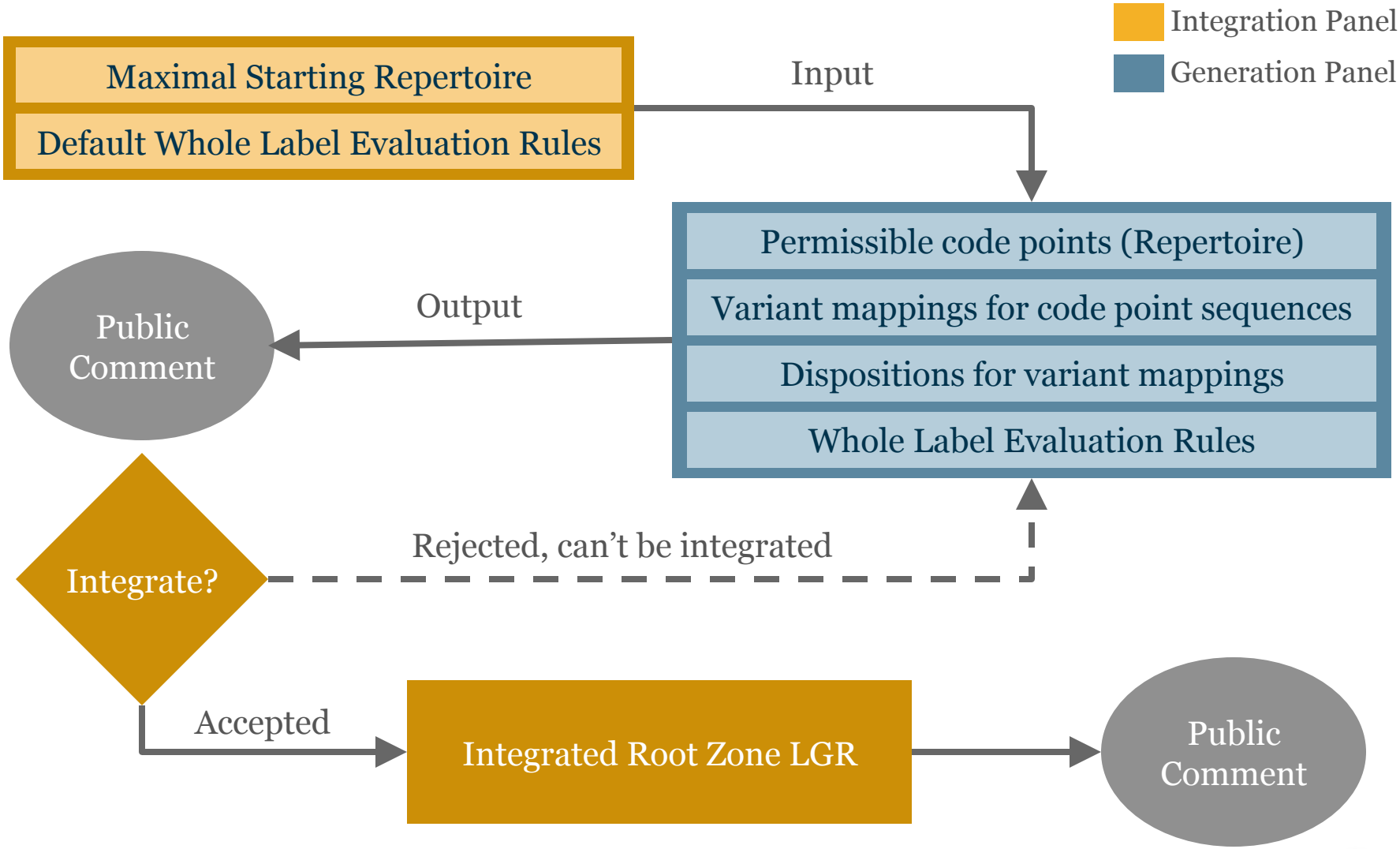
MSR-1 Status

- MSR-1: first Version of MSR
  - http://forum.icann.org/lists/comments-msr-03mar14/
- Released for Public Comment 3-March-2014
  - Comments close 21-May-2014
- Focuses on scripts for which IDN TLDs have been applied for
  - Also includes some closely related scripts
- Can be adjusted based on public comment
  - Will remain fixed after public review
- To be followed later by MSR-2 when more scripts are added

After MSR-1 Is Finalized

- Following the LGR Procedure:
  - Generation Panels select code points from MSR to include in their script LGRs' repertoire
  - Provide justification for inclusion
  - Add other elements of LGR
- Ongoing dialog between GP and IP before submission of LGR proposal
- IP will review and accept/reject script LGRs for integration

Flow for a Script-based LGR



MSR-1 content in numbers

- 22 scripts
  - Cyrillic, Georgian, Greek, Latin
  - Arabic, Hebrew
  - Han, Hangul, Hiragana, and Katakana
  - Bengali, Devanagari, Gujarati, Gurmukhi, Kannada, Malayalam, Oriya, Sinhala, Tamil, Telugu
  - Thai, Lao
- 'Common' and 'Inherited' (shared)
- 32,783 code points
  - 11,172 Hangul syllables and 19,849 Han ideographs

LGR: Modern Use in Everyday Writing

- What is a practical criterion for “modern use in everyday writing?”
- Many characters are only used in languages for smaller populations:
  - **Cyrillic:** smaller languages of the former Soviet Union
  - **Latin:** smaller languages of Africa, Pacific
- Goal: Selecting Code Points for the MSR by “Effective Demand”
  - Not merely size but language status and how it is used

Language Status From SIL Ethnologue

- EGIDS (Expanded Graded Intergenerational Disruption Scale)
  - Not based on population size, but on "established vitality"
- Used as proxy for "effective demand" for the writing system
  - Not a perfect correlation, some writing systems not stable
- For the MSR the IP used 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

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

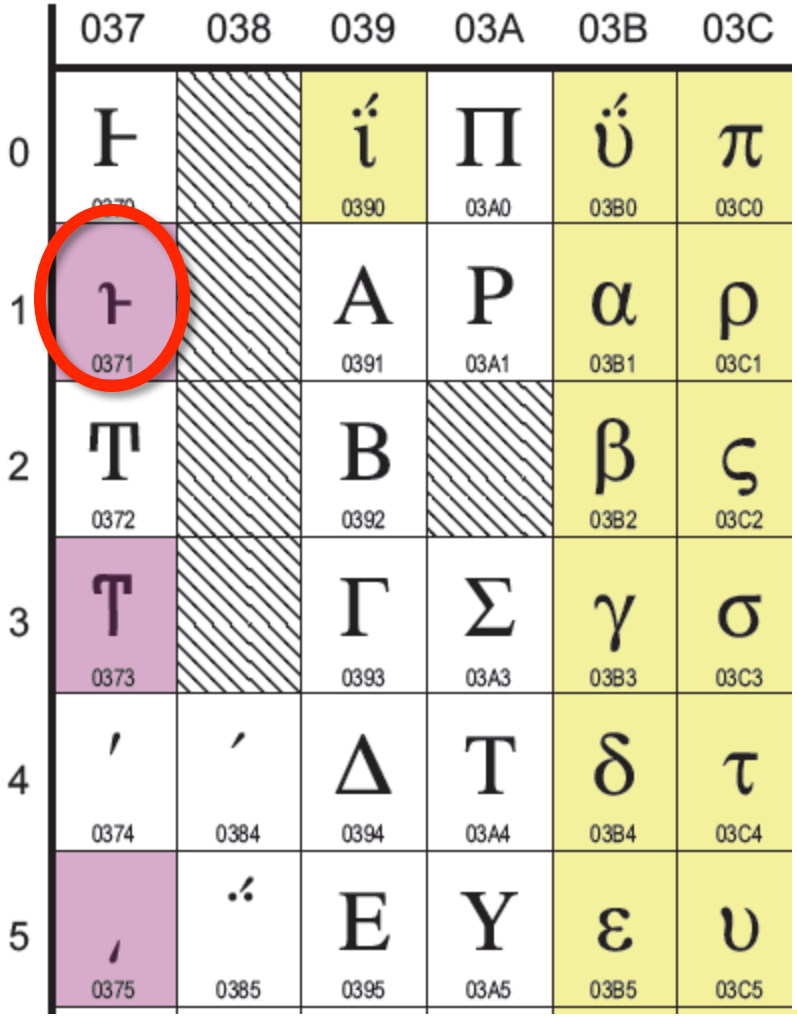
Code Points Excluded From the MSR

- Code points whose use is purely historic, or in limited or specialized use (for example in phonetic notation), are excluded from the root
- MSR excludes known historic, limited or specialized use code points that have no modern use in everyday writing
- Except:
  - Code points are allowed to remain in the MSR if their status could not be definitely confirmed and modern use in everyday writing could not be ruled out
    - For the LGR, Generation Panels are expected to confirm or exclude these code points

Reasons For Not Including Code Points

- Obsolete (historic, archaic)
- Limited or declining use (still developing, threatened, or nearly extinct)
- Symbol (characters classified as letters that are symbolic in nature)
- Punctuation (characters classified as letters that are punctuation-like)
- CONTEXTJ (Context — Joint controls)
- CONTEXTO (Context — Others)
- Unstable (encoding model changed)
- Deprecated (no longer in use, alternate code point sequence preferred)
- Technical use (phonetic, poetry)
- Religious use (annotation, cantillation)

Sample MSR Annotated Code Point Table



Archaic letters

0370 Ͱ GREEK CAPITAL LETTER HETA

0371 ͱ GREEK SMALL LETTER HETA
- obsolete

0372 Ͳ GREEK CAPITAL LETTER ARCHAIC SAMPI

0373 ͳ GREEK SMALL LETTER ARCHAIC SAMPI
- obsolete

Numeral signs

0374 ʹ GREEK NUMERAL SIGN

0375 ͵ GREEK LOWER NUMERAL SIGN
- ContextO
- obsolete

Archaic letters

0376 Ͷ GREEK CAPITAL LETTER PAMPHYLIAN DIGAMMA

0377 ͷ GREEK SMALL LETTER PAMPHYLIAN DIGAM

- Not PVALID
- In MSR-1
- Excluded from MSR-1

Combining Sequences

- Sequence of base character and combining mark
- Some sequences also exist in pre-composed form
- IDNA 2008 requires use of pre-composed form (NFC)
- Some writing systems use a number of sequences beyond pre-composed characters (example Latin)
- Some scripts use combining marks extensively, but optionally, (e.g., as vowels in Arabic)
- They are an essential part of other Scripts (e.g., used for dependent vowel signs in Indic)

Combining Sequences

- Combining marks in Block [0300-036F]
  - Mainly in conjunction with Latin, Greek and Cyrillic
- Writing systems for some languages in these scripts require combining sequences
- Greek
  - All monotonic combinations included, using tonos and dialytika (0301, 0308 and their combination 0344)
  - No need for combining marks in Greek repertoire

Latin and Cyrillic

- Writing systems using Latin and Cyrillic scripts can express most combinations in pre-composed forms
  - Few exceptions, such as U+0329 VERTICAL BAR BELOW in Yoruba
- MSR includes any *required* combining characters which do not appear in pre-composed characters
- MSR also includes combining marks that are part of pre-composed characters, in case they have further uses
- The Generation Panels are expected to
  - Investigate and only include those combining marks for which additional, not pre-composed combinations are documented
  - Consider restricting combining marks to a fixed set of combinations of base + combining character where feasible

Options for Other Scripts

- Consider excluding altogether
  - in scripts used for languages where omission is usual
  - where marks mimic punctuation
  - where fidelity of display is too poor to distinguish them
- Consider listing a fixed set of combinations:
  - where few combinations are ever used in practice
- Consider restriction to well formed clusters
  - may imply whole label rules (whole aksharas for Indic)
- Consider full use on a par with other code points

Han Characters in MSR

- Han ideographs have no hard division into “common, modern, every-day” and “historic, archaic, limited-use”
- IP used following heuristic
  - Superset of existing IDN tables for Han ideographs
  - IICORE repertoire
- Result will be outer limit of ideographs for use in Root Zone, once MSR is finalized
  - Public Review allows to include missing code points

Han Characters in MSR-1

Annotated code point tables for Han characters give additional information about the code point that is solely intended to facilitate review.

What About Variants?

- Code point variants are not defined in the MSR
  - Examples*: U+4C81:鰯 ↔U+9CDA 鰯 or Strasse, Straße
- Identifying code point variants and assigning dispositions is a task for Generation Panels
  - Dispositions: ***blocked*** or ***allocatable***
- Integration Panel will look for:
  - Justification for variants
  - Minimizing use of allocatable variants per conservatism principle
  - Preferred use of blocked variants to prevent issues

* These are conceptual examples, not suggestions

Whole Label Evaluation Rules (WLE)

- Evaluate code points in context of whole label
  - Prevent labels that cannot be reliably displayed
    - Restrict Indic labels to sequences of well-formed syllables
- Might also be used to block a variant label, even if all code points in it are otherwise valid
  - Method to limit allocatable variants
- MSR-1 contains a single default WLE rule
  - Prevents leading combining marks

XML Format

- Normative definition of MSR-1 in XML format <link to internet draft>
- Format is not specific to Root Zone LGR
  - Not all features will be used
- Simple script-LGR can just delete unneeded code points from MSR-1 and update meta data, such as date, description, etc.
- For complex cases, IP and ICANN will assist
  - Conversion from IDN table formats can be scripted

Session 3: Training on XML format to represent the LGR

A Brief Tutorial

What This Tutorial Is About?

- Goals for the new format
- Main processing steps that can be performed
- Elements of the LGR
  - Code Points and Variants
  - Character Classes
  - Rules
  - Actions
  - Default Actions

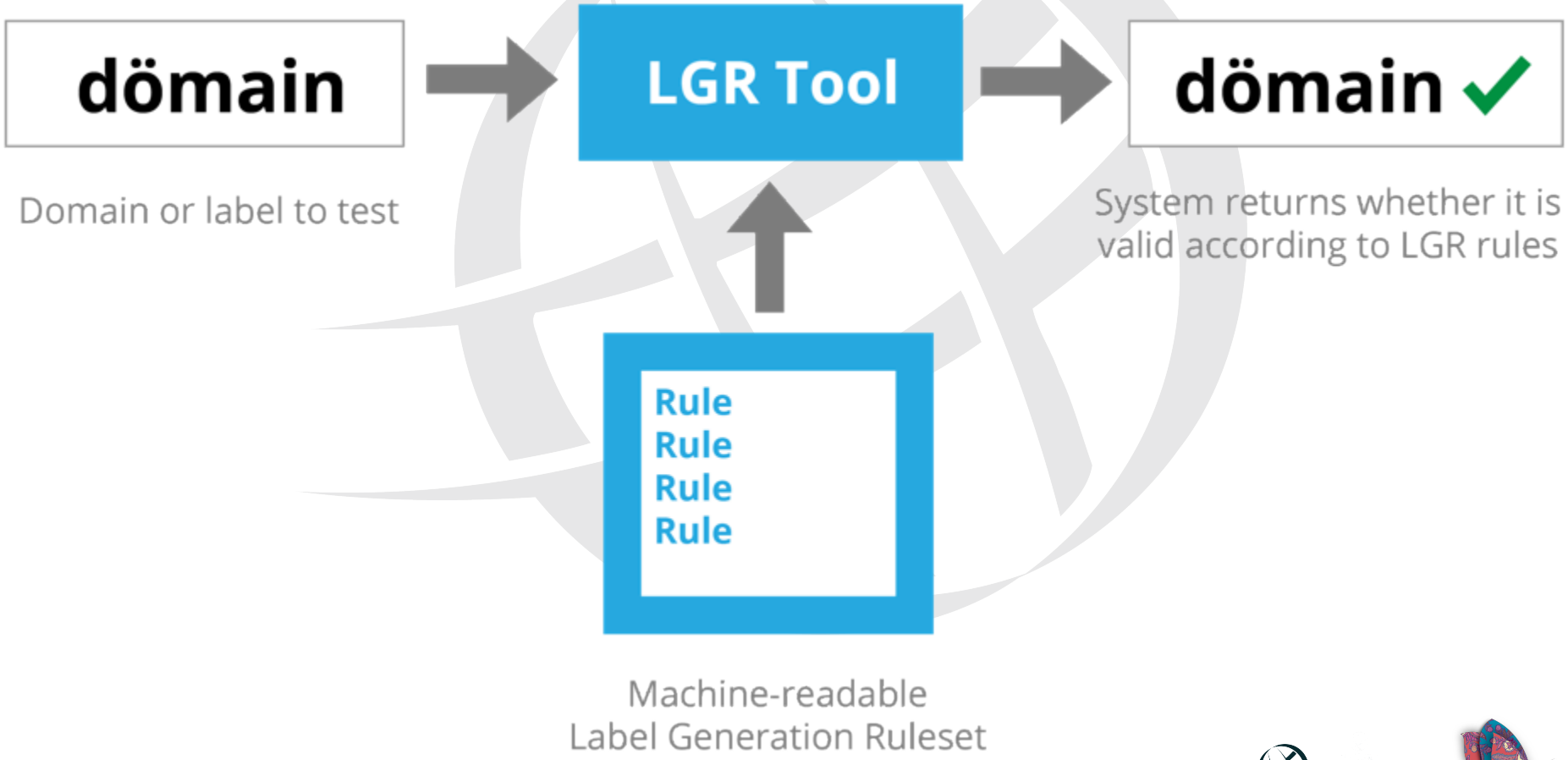
Basic Goals

- Label Generation Rulesets define permitted labels and variants (aka IDN Tables)
- Published by a registry as part of its policies, and used to check validity of labels submitted for registration
- XML representation:
  - Common format
  - Machine Readable
  - Can represent existing IDN Tables
  - Make certain policy rules explicit
  - Standard format for Root Zone LGR*

* Root Zone LGR will not use all features of the XML format

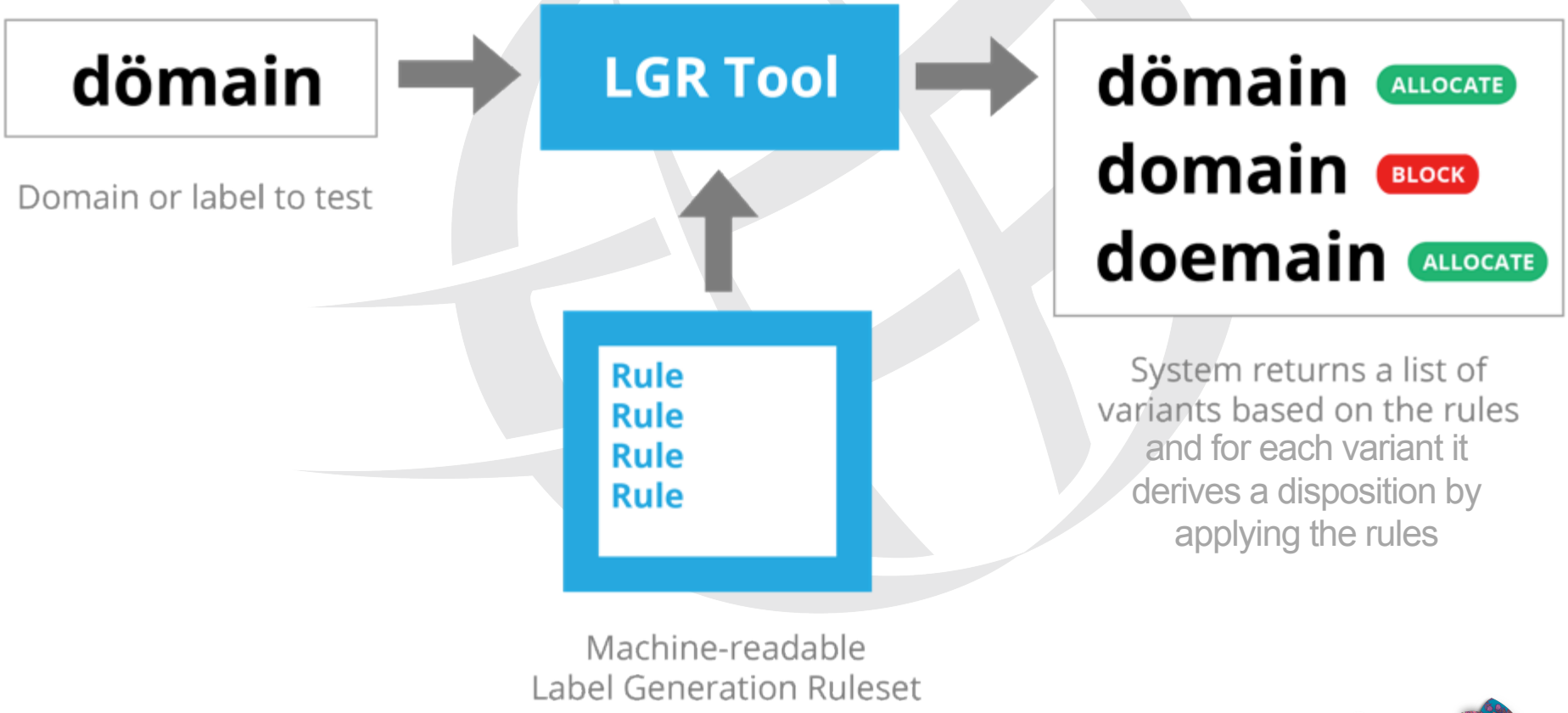
What Does XML-LGR Enable?

Simple Validity Checking

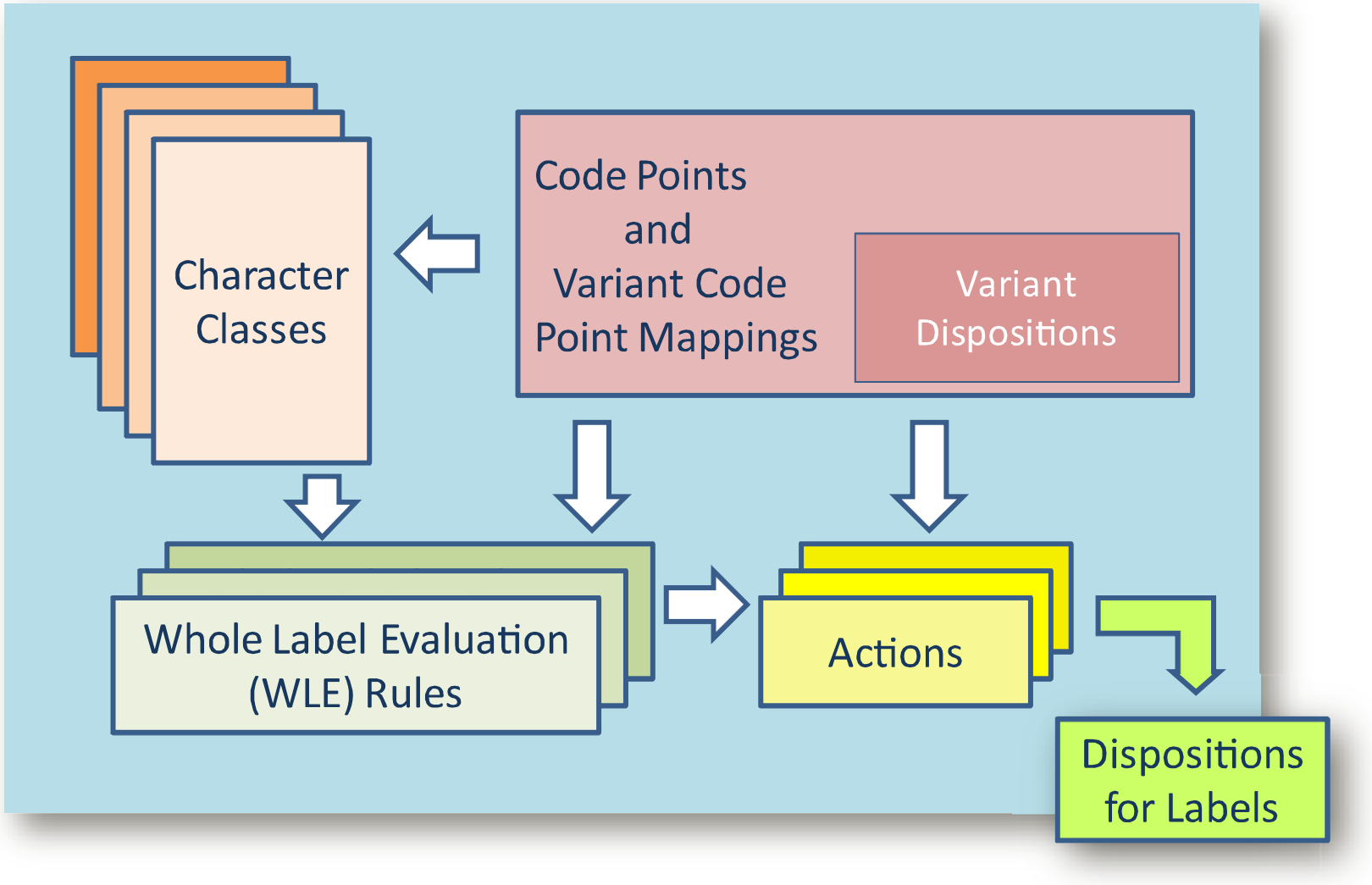


What Does XML-LGR Enable?

Variant Label Generation and Disposition



What Does an LGR Contain?



Main Elements of an LGR

- Code point list (mandatory)
  - Specific code points that are allowed or disallowed
  - Includes optional variant mappings to other code points
- Derived code point sets (optional)
  - Explicitly listed, or based on Unicode properties, such as Greek script characters, or those with specific joining properties
- Contextual "whole-label" rules (optional)
  - Results depend on where characters appear in the label relative to other characters,
    used to evaluate "whole labels"
  - Can be used to disallow code point or variant in some context
- Actions (use default if not supplied)
  - Define actions to take based on a given rule or variant dispositions: allocate, block, etc.

The XML File

```
<?xml version="1.0" encoding="utf-8"?>
<lgr xmlns="http://www.iana.org/lgr/0.1">
 <meta>
        Version, description, domain, etc…
 </meta>
 <data>
        Code points and variants…
 </data>
 <rules>
        Character classes, Rules and Actions…
 </rules>
</lgr>
```

<meta> : “Properties” for the File

```
<meta>
        <version>1</version>
        <description type=“text/plain”>
         this is an example
        </description>
        <language>und-Latn</language>
        <domain>.example</domain>
</meta>
```

- Use <language> element with “und-” for script-specific LGR

<data>: Code Point Repertoire

- Define code points eligible in labels:
  - Code point

```
<char cp="002D" />
```

  - Range

```
<range first-cp="0030" last-cp="0039"/>
```

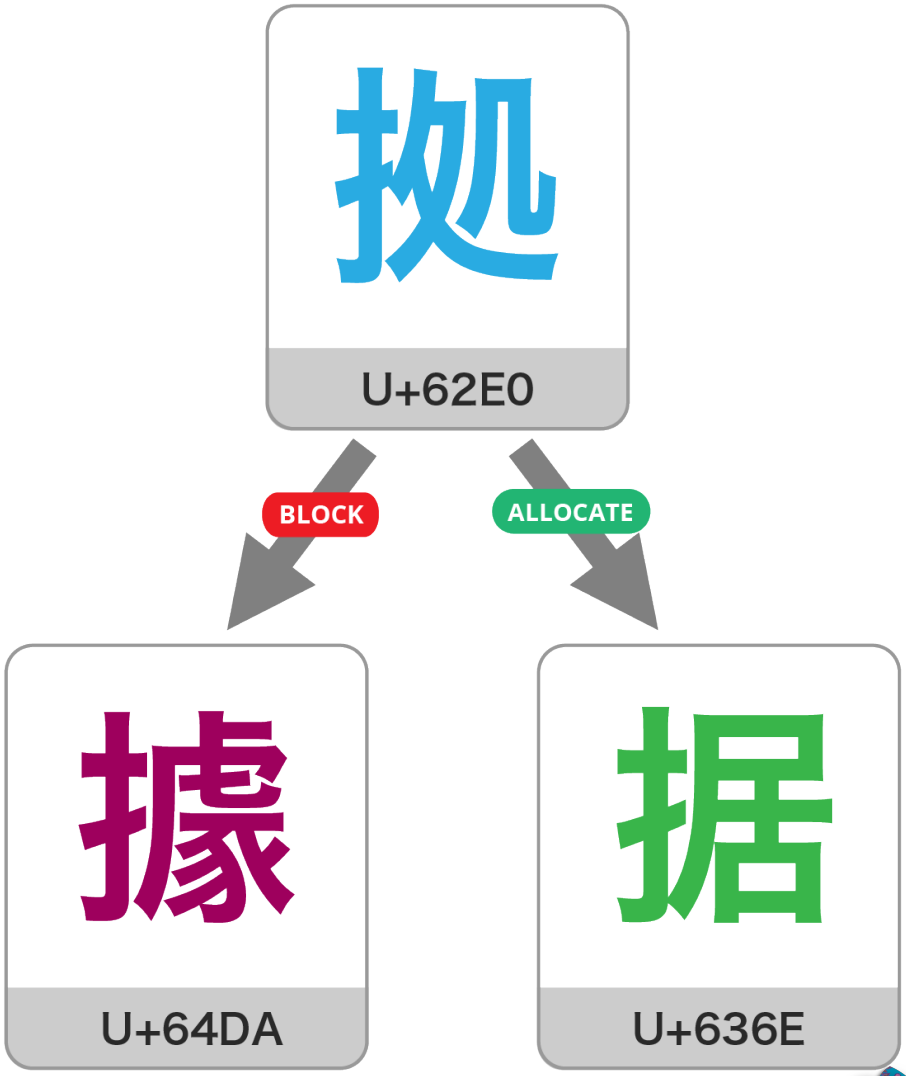
  - Sequence

```
<char cp="006C 00B7 006C">
```

- Code Points and Sequences may have variants

<data>: Code Point Variant Example

```
<char cp="62E0">
   <var cp="636E" disp="allocate"/>
   <var cp="64DA" disp="block"/>
</char>
```

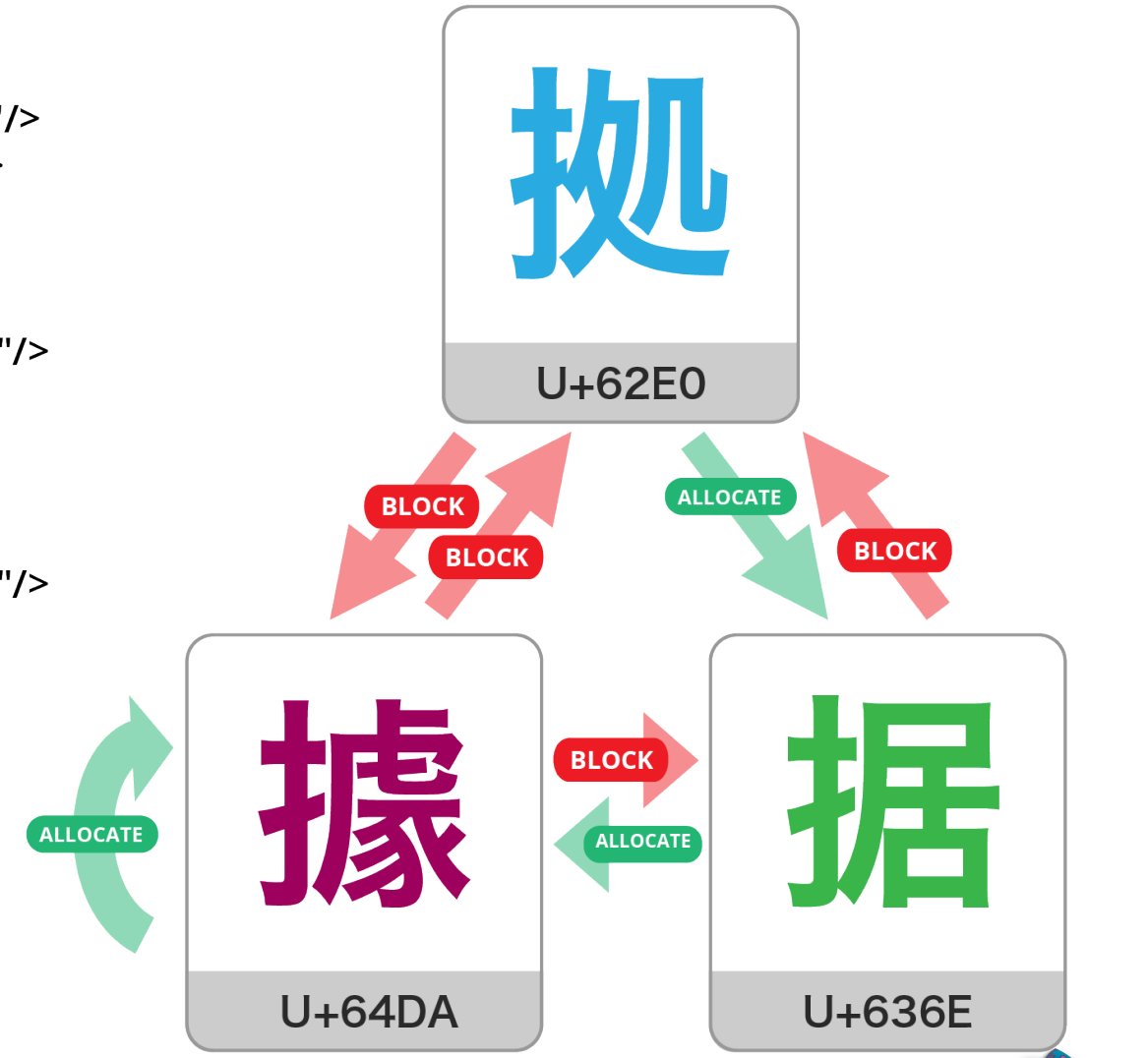


- Chinese characters can have simplified or traditional forms. These are defined as variants of each other
- Variants have dispositions that define whether variant labels should be allocated or blocked

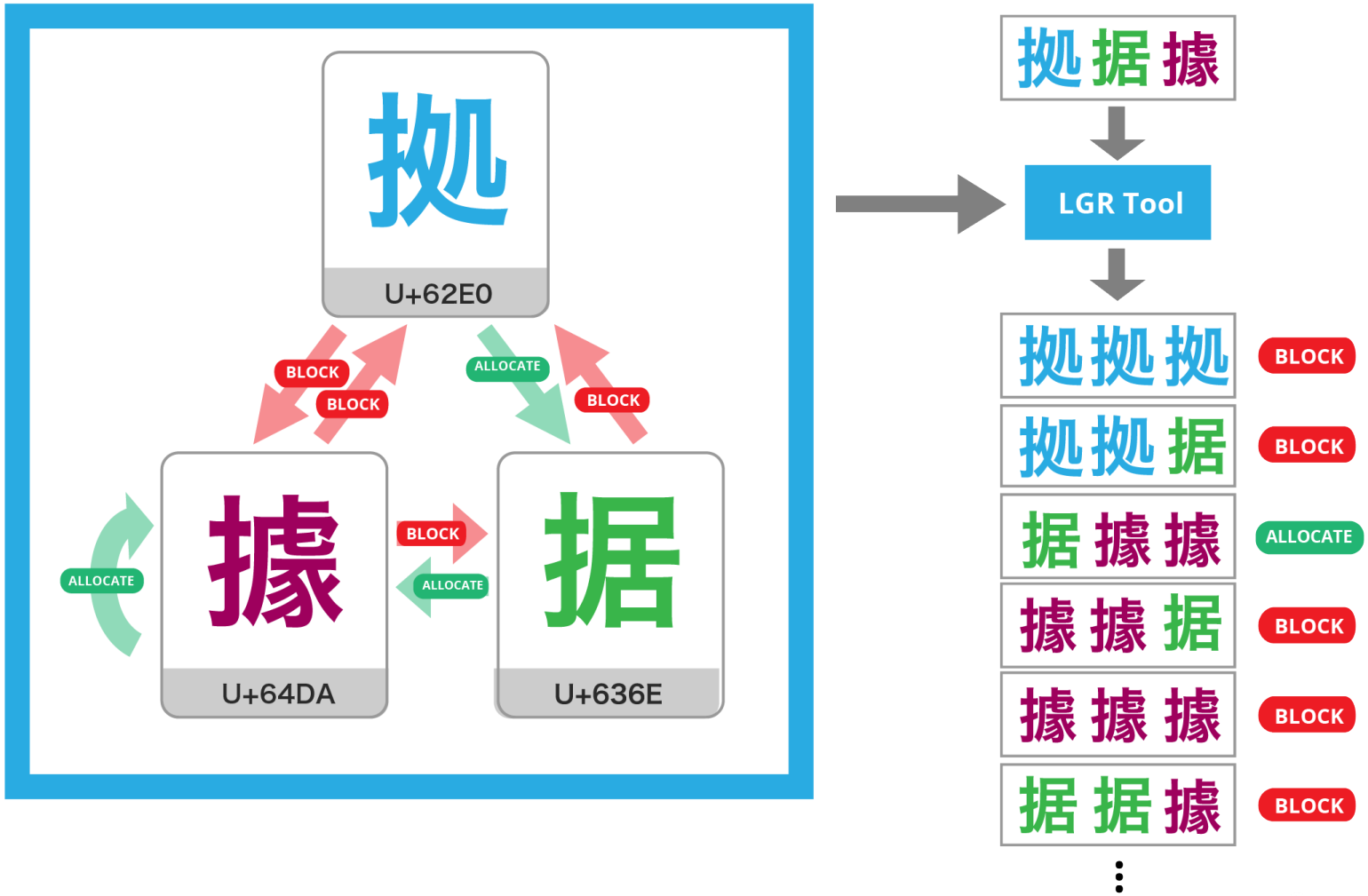
Variants Example, Fully Defined

```
<char cp="62E0">
   <var cp="636E" disp="allocate"/>
   <var cp="64DA" disp="block"/>
</char>
<char cp="636E">
   <var cp="62E0" disp="block"/>
   <var cp="64DA" disp="allocate"/>
</char>
<char cp="64DA">
   <var cp="62E0" disp="block"/>
   <var cp="636E" disp="block"/>
   <var cp="64DA" disp="allocate"/>
</char>
```

Variant relationships are symmetric and transitive and may including identity mapping



From Variant Definitions To Label Dispositions



<rules>: Classes, Rules and Actions

- Classes, rules and actions define
  - Which labels are valid
  - Whether labels are well-formed
  - How to derive the disposition of variant labels from disposition for variants
    - Supports RFC 3743 rules for variants
- Default actions for common cases
- Minimal “built-in” knowledge
  - IDNA 2008 rules and all typical policy rules can be made explicit for machine evaluation

Whole Label Evaluation (WLE) Rules

- Rules are similar to regular expressions
- They match certain labels or variants
- Example of a WLE rule matching labels that start with a non-spacing mark (e.g., accent)*

```
<rule name="leading-nonspacing-mark">
   <start/>
   <class property="gc:Mn"/>
</rule>
```

* General category value Mn = non-spacing marks (accents)

Used in Rules: Character Classes

Sets of code points defined by:

- Explicit list of code points

```
<class name="digits">0030-0039</class>
```

- Unicode property

```
<class property="jt:D" />
```

- "Tag" attributes on code points

```
<char cp="4E00" tag="preferred" />
```

- Set Operator

```
<complement>
  <class from-tag="preferred" />
</complement>
```

Used in Rules: Match Operators

- The following match operators are supported

  <**start**>, <**end**>
  <**any**>, <**char**>
  <**class**>
  <**choice**>, <**rule**> (for grouping)
  <**look-behind**>, <**look-ahead**>, <**anchor**>

- Equivalents to regular expressions

  <**start**> = ^
  <**any** count="0+" /> = .*
  <**class**>0030-0039</**class**> = [0-9]

When Rules: Apply in Context

- Variants or code points may require context
- Code point 200D is invalid unless except:

```
<char cp="200D" when="follows-virama" />
```

- Rule matches when 200D is preceded by a virama*
  - <anchor/> represents location of 200D in context

```
<rule name="follows-virama">
  <look-behind>
    <class property="ccc:9" comment="virama" />
  </look-behind>
  <anchor />
</rule>
```

Actions: Assign Dispositions

- Actions set dispositions for labels and variant labels
- Executed in order; first triggered action counts
- Example of an action triggered by the label matching a rule and setting the label disposition to "invalid"

```
<action
        disp="invalid"
        match="leading-nonspacing-mark"
/>
```

- Other types of triggers defined

What Triggers an Action?

Actions are triggered:

- When a *rule* is matched or not matched
- When a label contains *any variant* with a given disposition
- When a label *exclusively contains variants* of a given disposition
- When *all variants* in a label are of a given disposition, but may include code points without variant mappings

Some Default Actions

- Block any variant label with a single blocking variant code point

  <action disp="block" any-variant="block" />

- Mark as allocatable any variant label where all code points are either allocatable variant code points or original code points

  <action disp="allocate" all-variants="allocate" />

- Mark as allocatable any valid label (catch-all)*

  <action disp="allocate" />

* Occurs last in order of precedence

In Summary:
Label Generation Rulesets

- Expand the notion of IDN tables to beyond just simple code point lists
- Provide comprehensive way to describe registry policies relating to permitted code points
- Universal format that can be implemented in a single fashion across multiple domains and policies
- Good basis for clear and consistent rulesets for the root zone

Resources

- Working with XML
  - xmllint —relaxng /path/to/kjd/lgr/specification/idn-tables-1.0.rng —noout MSR-1-Repertoire+WLE-Rules.xml
  - RelaxNG Schema for validating
  - https://github.com/kjd/lgr/blob/master/specification/idn-tables-1.0.rnc
- Further resources at github.com/kjd/lgr
  - Internet draft
  - Correspondence to regex
  - IDNA 2008 rules in XML format

Thank You

USEFUL LINKS:

- The LGR Procedure: http://www.icann.org/en/resources/idn/variant-tlds/lgr-procedure-20mar13-en.pdf
- Call for Generation Panels to Develop Root Zone Label Generation Rules: http://www.icann.org/en/news/announcements/announcement-11jul13-en.htm
- MSR Version 1 Available for Public Comment: http://www.icann.org/en/news/public-comment/msr-03mar14-en.htm
- V06 Internet Draft for LGR Rules Toolset Project Published: http://tools.ietf.org/html/draft-davies-idntables
- Generation Panel for Arabic Script Root Zone Label Generation Rules Seated: http://www.icann.org/en/news/announcements/announcement-2-14feb14-en.htm
- Selection of Integration Panel for the IDN Root Zone Label Generation Rules: http://www.icann.org/en/news/announcements/announcement-06sep13-en.htm
- Setting up and running a Generation Panel: https://community.icann.org/display/croscomlgrprocedure/Generation+Panels
- Community Wiki LGR Project website: https://community.icann.org/display/croscomlgrprocedure/Root+Zone+LGR+Project
- Examining the User Experience Implications of Active Variant TLDs: http://www.icann.org/en/resources/idn/variant-tlds/active-ux-21mar13-en.pdf