ICANN|55 MARRAKECH

5 – 10 MARCH 2016

ICANN

IDN Root Zone LGR Workshop

Sarmad Hussain | ICANN 55 | 9 March 2016

Agenda

⊙ LGR Toolset	– Marc Blanchet
⊙ LGR Considerations	– Michel Suignard, IP
⊙ Community Updates	
• Thai GP	– Pitinan Kooarmornpatana
• Korean GP	– KIM Kyongsok
• Japanese GP	– Hiro Hotta
• Chinese GP	– Wang Wei
⊙ Q/A	

LGR Toolset

Marc Blanchet | IDN LGR Workshop | March 9 2016

Agenda 1 Slide

1

Project

2

New Features

3

Conclusion

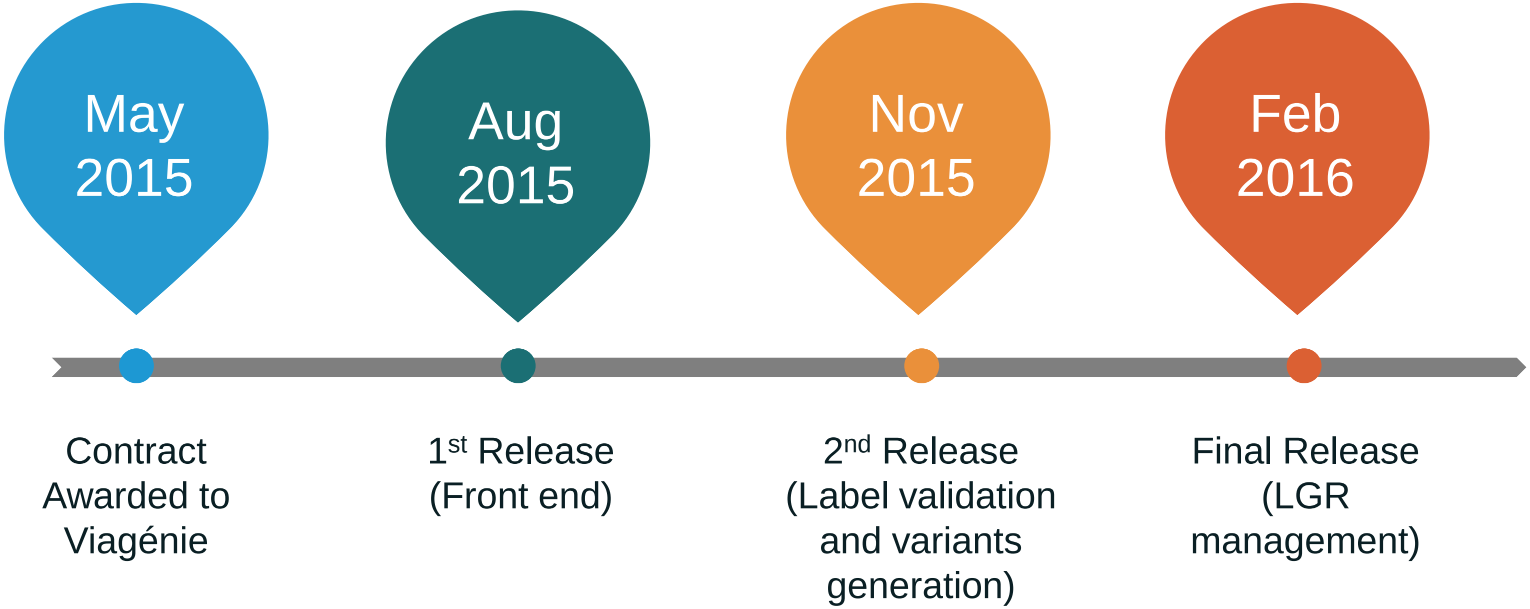
Project

Project

- Web-based LGR editing tool.
- 3 Phases:
  - 1. LGR creation: Web-based application.
  - 2. Select a pre-defined LGR in the XML Format, and validate a label or generate its variant labels along with their dispositions.
  - 3. LGR management functions: conversion of language tables into the XML Format, comparing two LGRs, and additional operations including union, intersection and difference of two LGRs.

  - Implementation team: Julien Bernard, Marc Blanchet, David Drouin, Audric Schiltknecht, Wil Tan.

Timeline



May 2015

Contract Awarded to Viagénie

Aug 2015

1st Release (Front end)

Nov 2015

2nd Release (Label validation and variants generation)

Feb 2016

Final Release (LGR management)

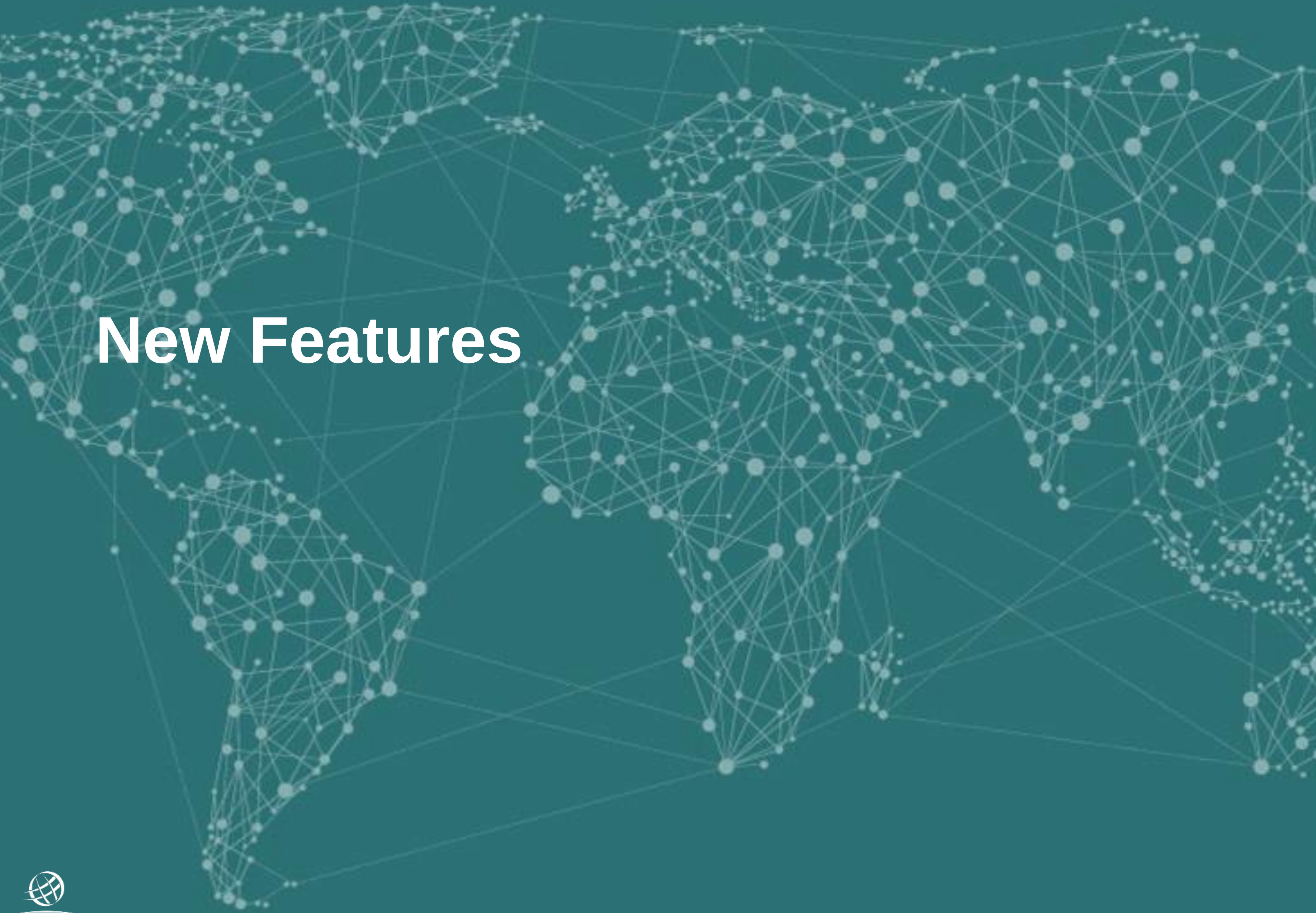
Updates Since Phase 2

- WLE rules edition
  - Edit/define/delete new classes, rules and actions.
  - Define tags in code point view to be referenced in classes.
- Tag edition on code points.
- LGR management: Comparison of LGRs (union/intersection), diff.
- Button to expand ranges to single code points in repertoire.
- Button to save LGR summary.
- Add variants when importing code points from file in manual mode.
- Update namespace: Automatically convert old LGR files to use new namespace.
- Security audit enhancement.
- Docker-ready.

- Available on lgr-demo.viagenie.ca
- Available on ICANN web server: https://lgrbuilder.icann.org/

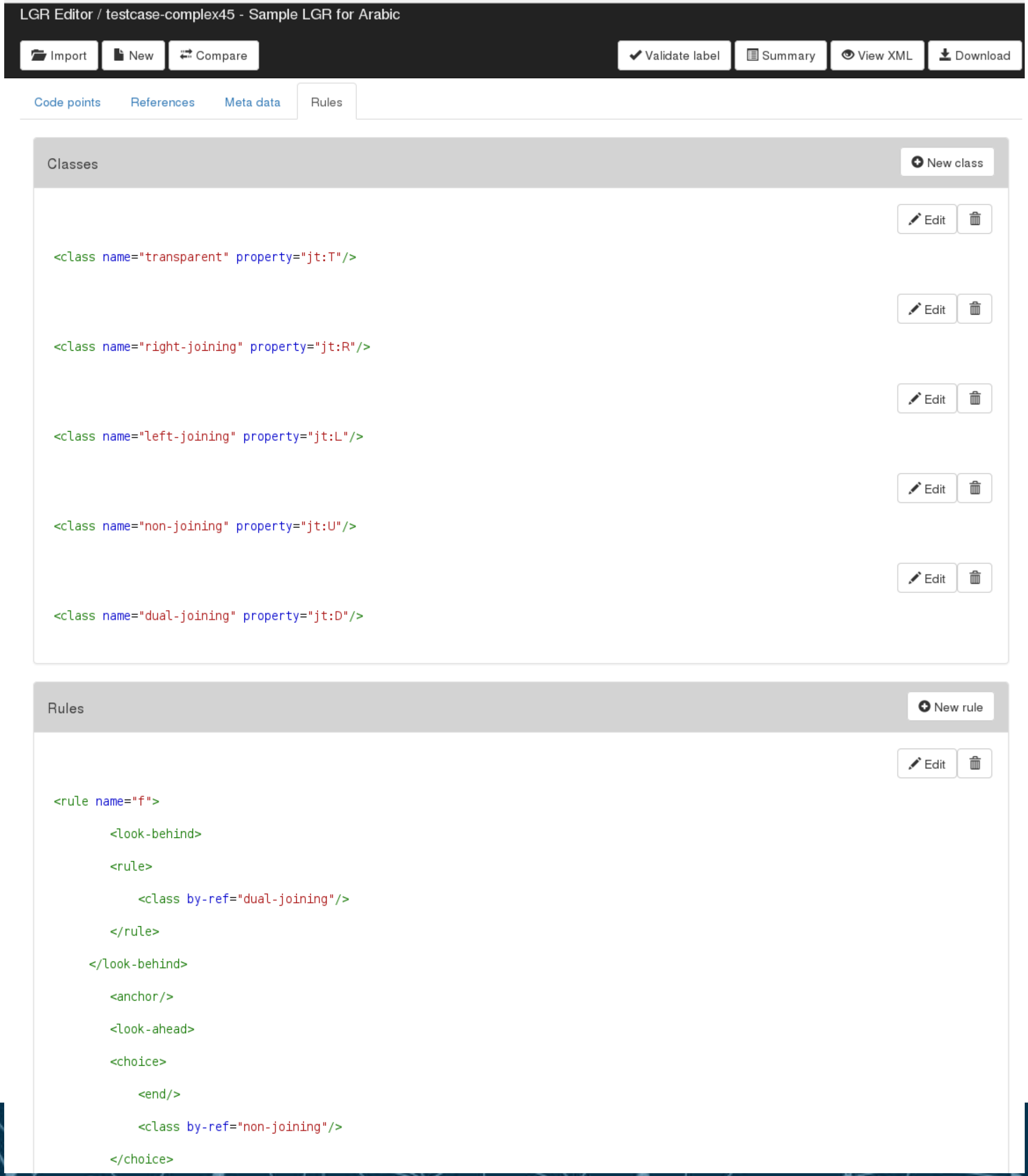
On Unicode Version

- Editor is Unicode version independent, but applies IDNA2008 rules.
- However, any LGR processing depends on specific Unicode version.
- Tool is based on Unicode 6.3, aligned with current IDNA IANA tables.



New Features

WLE Rules Edition - List Rules



```
<class name="transparent" property="jt:T"/>

<class name="right-joining" property="jt:R"/>

<class name="left-joining" property="jt:L"/>

<class name="non-joining" property="jt:U"/>

<class name="dual-joining" property="jt:D"/>
```

```
<rule name="f">
        <look-behind>
        <rule>
            <class by-ref="dual-joining"/>
        </rule>
     </look-behind>
        <anchor/>
        <look-ahead>
        <choice>
            <end/>
            <class by-ref="non-joining"/>
        </choice>
```

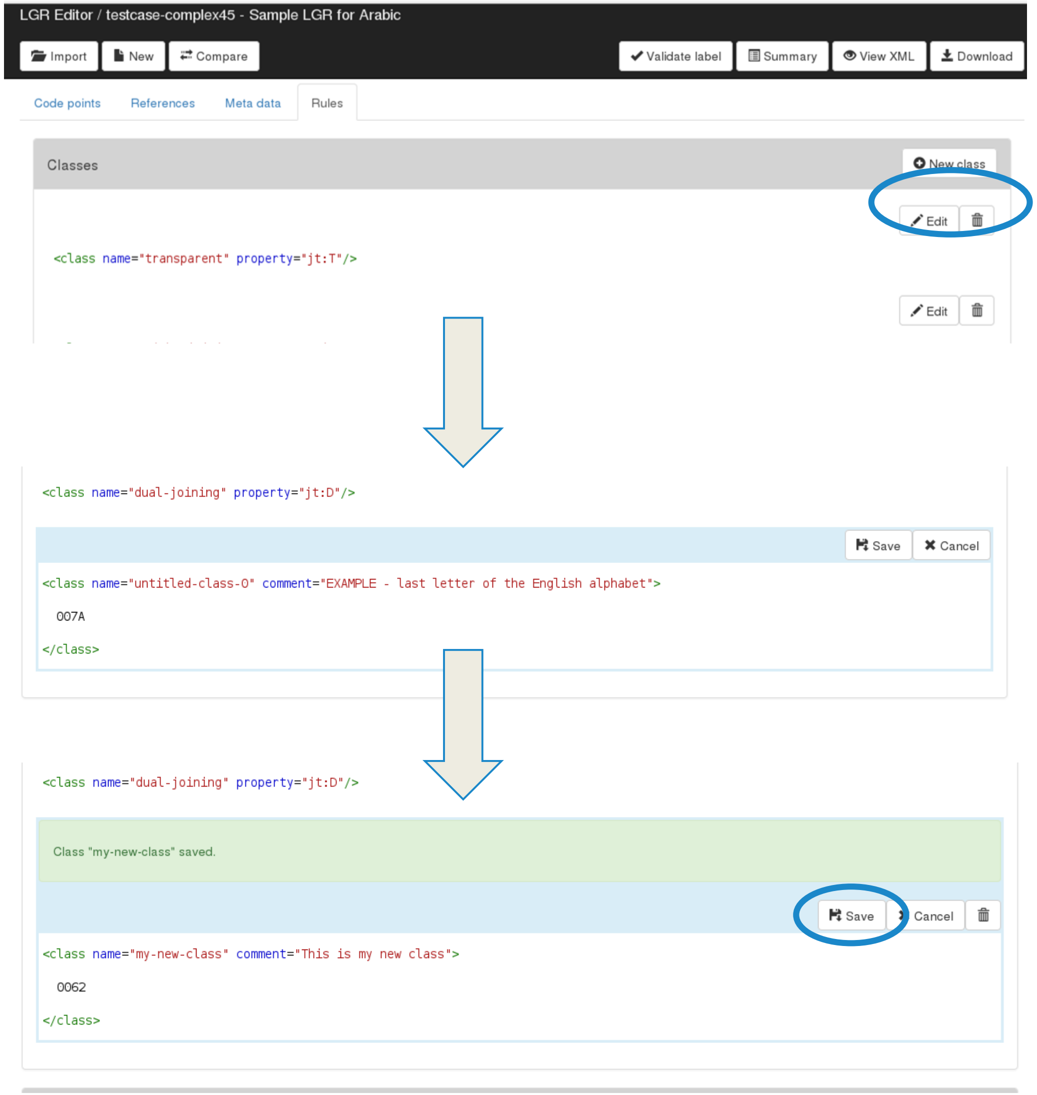
WLE Rules Edition - Add Class

- Tool provides a template to edit when adding a new.

- When saving, basic validation is done.

- Advanced and full LGR validation is done when doing the “Summary”.

- Same for Classes, Rules and Actions.



```
<class name="transparent" property="jt:T"/>
```

```
<class name="dual-joining" property="jt:D"/>
```

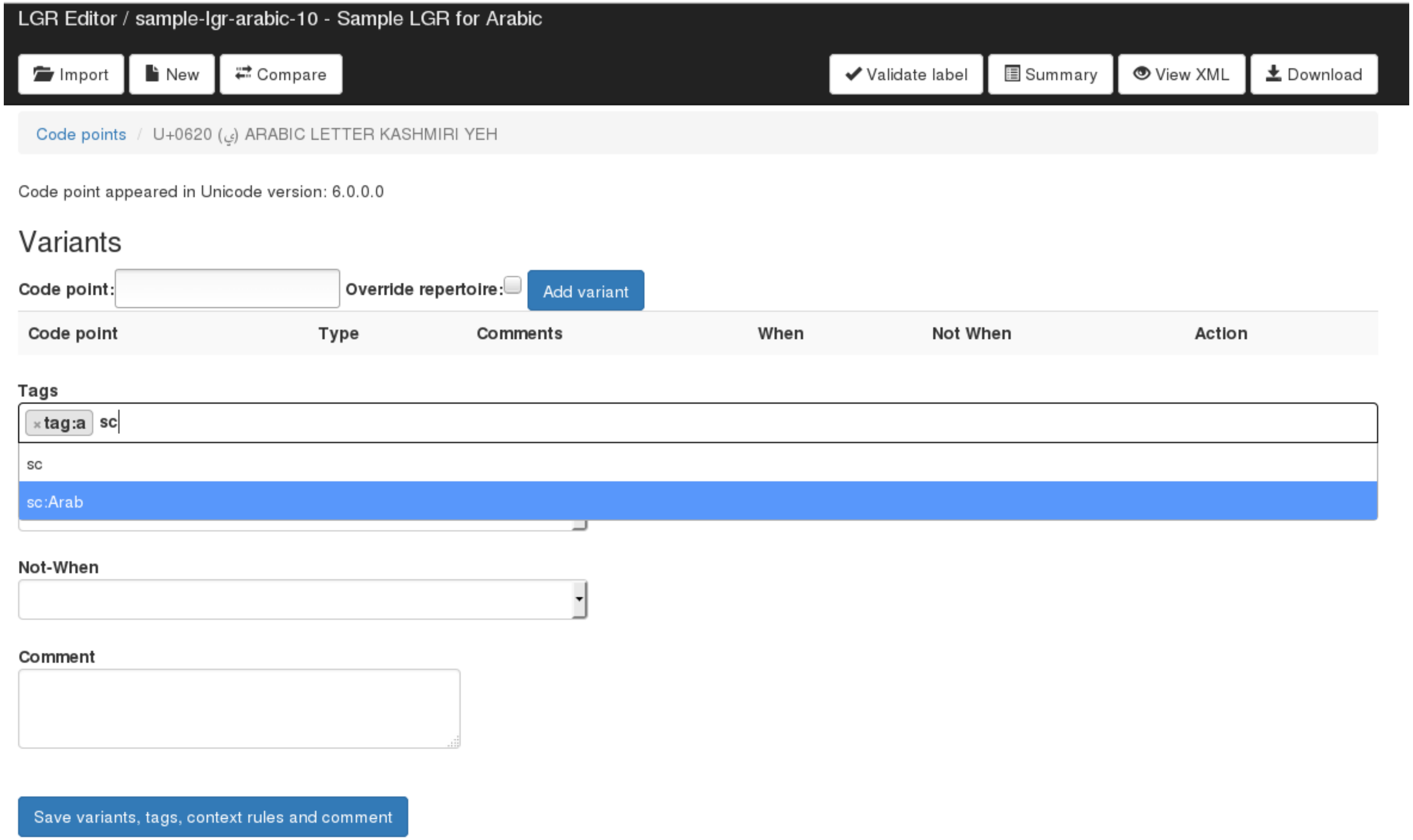
```
<class name="untitled-class-0" comment="EXAMPLE - last letter of the English alphabet">
  007A
</class>
```

```
<class name="dual-joining" property="jt:D"/>
```

Class "my-new-class" saved.

```
<class name="my-new-class" comment="This is my new class">
  0062
</class>
```

Tag Edition (autocompletion)



LGR Editor / sample-lgr-arabic-10 - Sample LGR for Arabic
Import
New
Compare
✔ Validate label
Summary
View XML
Download
Code points / U+0620 (ؠ) ARABIC LETTER KASHMIRI YEH

Code point appeared in Unicode version: 6.0.0.0

Variants

Code point: ____ **Override repertoire:** □ Add variant

Code point	Type	Comments	When	Not When	Action

Tags

× tag:a sc

sc

sc:Arab

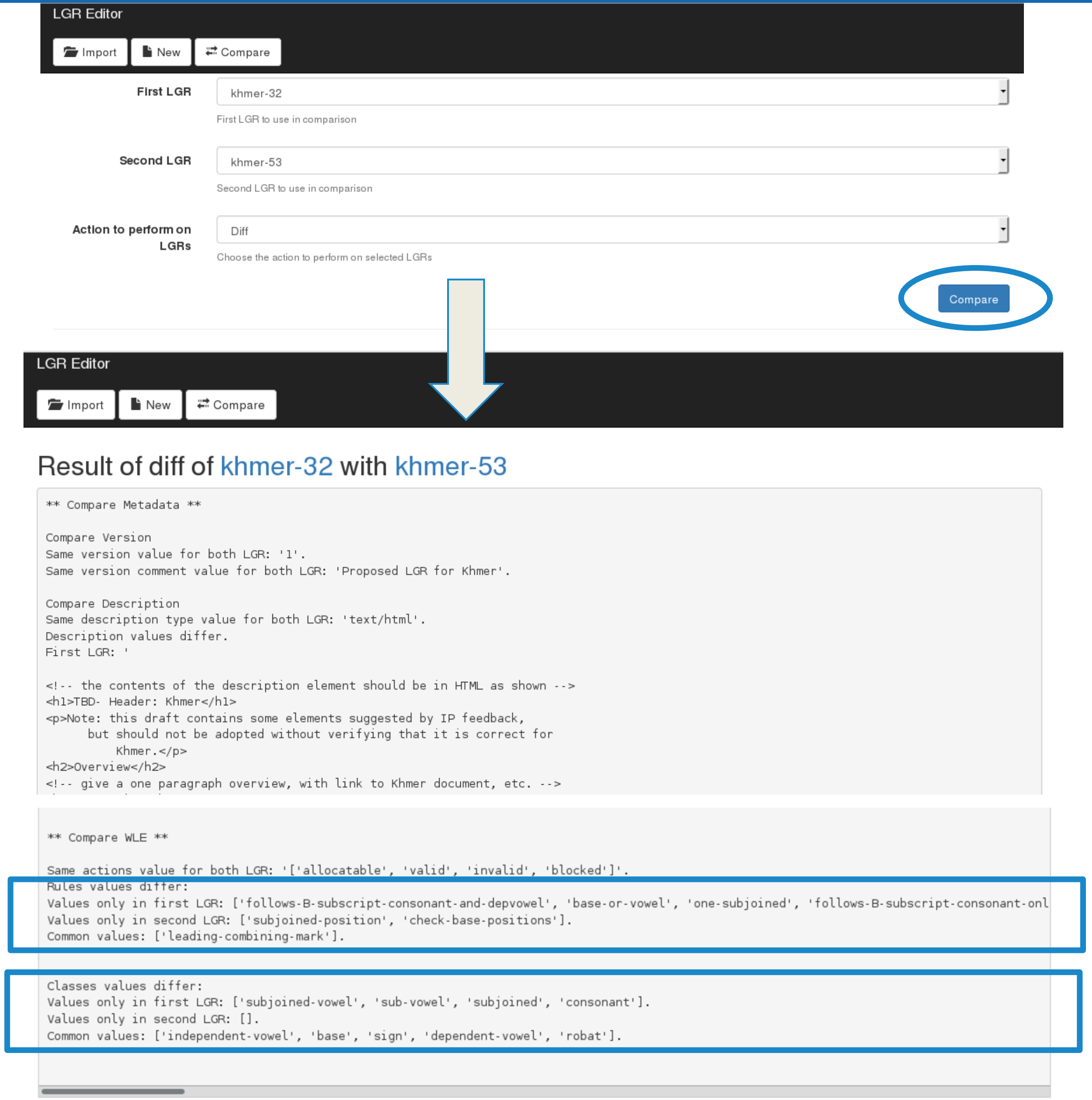
Not-When

Comment

Save variants, tags, context rules and comment

LGR Diff

- Input are 2 LGRs in the workspace.
- Action: Union, Intersection or Diff.
- Diff shown.



First LGR: khmer-32
First LGR to use in comparison

Second LGR: khmer-53
Second LGR to use in comparison

Action to perform on LGRs: Diff
Choose the action to perform on selected LGRs

Result of diff of khmer-32 with khmer-53

```
** Compare Metadata **

Compare Version
Same version value for both LGR: '1'.
Same version comment value for both LGR: 'Proposed LGR for Khmer'.

Compare Description
Same description type value for both LGR: 'text/html'.
Description values differ.
First LGR: '

<!-- the contents of the description element should be in HTML as shown -->
<h1>TBD- Header: Khmer</h1>
<p>Note: this draft contains some elements suggested by IP feedback,
      but should not be adopted without verifying that it is correct for
          Khmer.</p>
<h2>Overview</h2>
<!-- give a one paragraph overview, with link to Khmer document, etc. -->
```

```
** Compare WLE **

Same actions value for both LGR: '['allocatable', 'valid', 'invalid', 'blocked']'.
Rules values differ:
Values only in first LGR: ['follows-B-subscript-consonant-and-depvowel', 'base-or-vowel', 'one-subjoined', 'follows-B-subscript-consonant-onl
Values only in second LGR: ['subjoined-position', 'check-base-positions'].
Common values: ['leading-combining-mark'].

Classes values differ:
Values only in first LGR: ['subjoined-vowel', 'sub-vowel', 'subjoined', 'consonant'].
Values only in second LGR: [].
Common values: ['independent-vowel', 'base', 'sign', 'dependent-vowel', 'robat'].
```

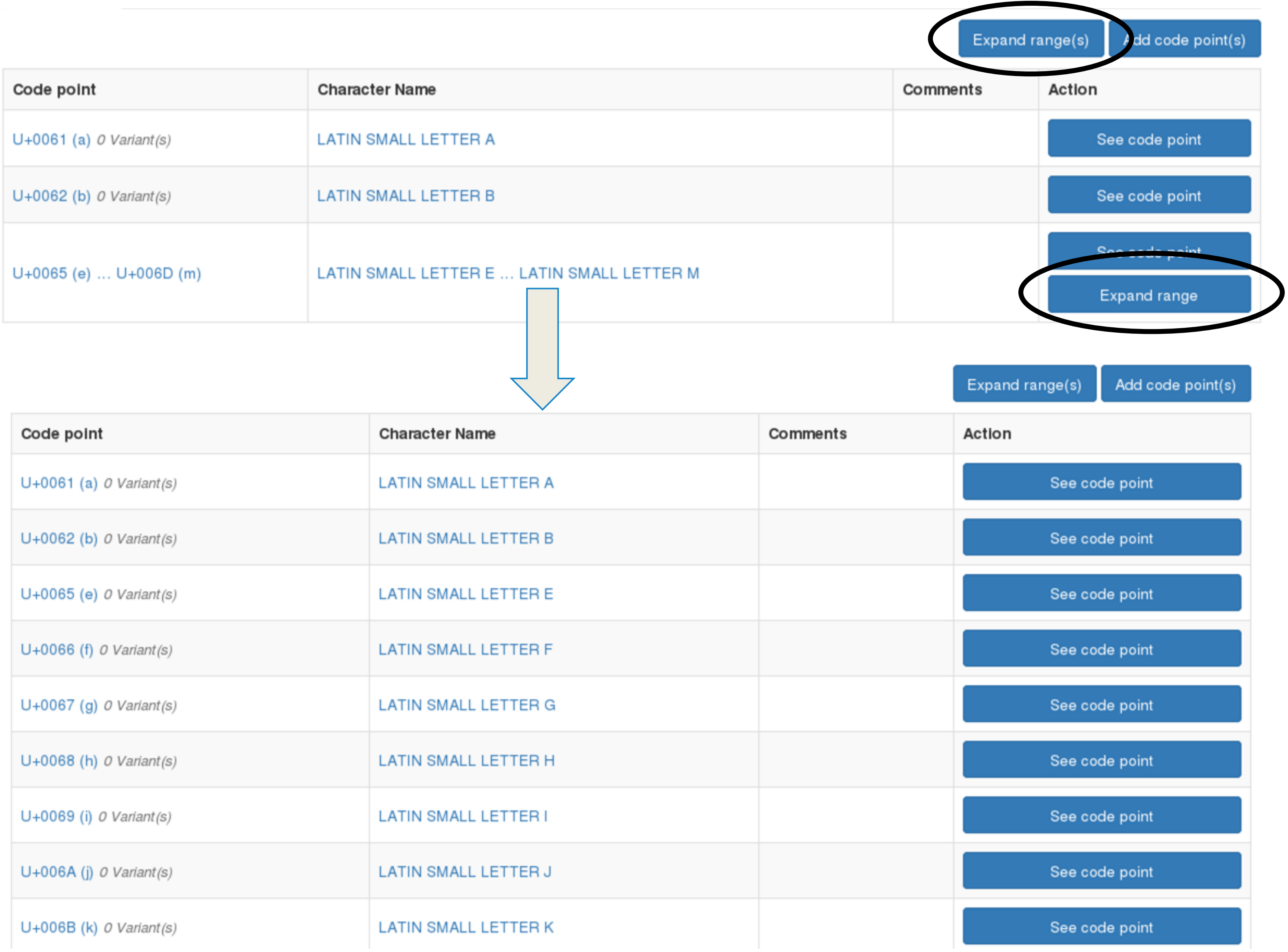
LGR Union and Intersection

- Union and Intersection:
  - Result creates a new LGR.
  - Put into the user’s workspace.
  - Directly displayed in the code point view.
  - Name of the resulting LGR is automatically created.

Expansion of Ranges

- Convert the range into its list of codepoints.
- Convert a specific range or all ranges.
- Generated codepoints inherit properties of the range (i.e. tags, comments, references, …)

Expansion of Ranges

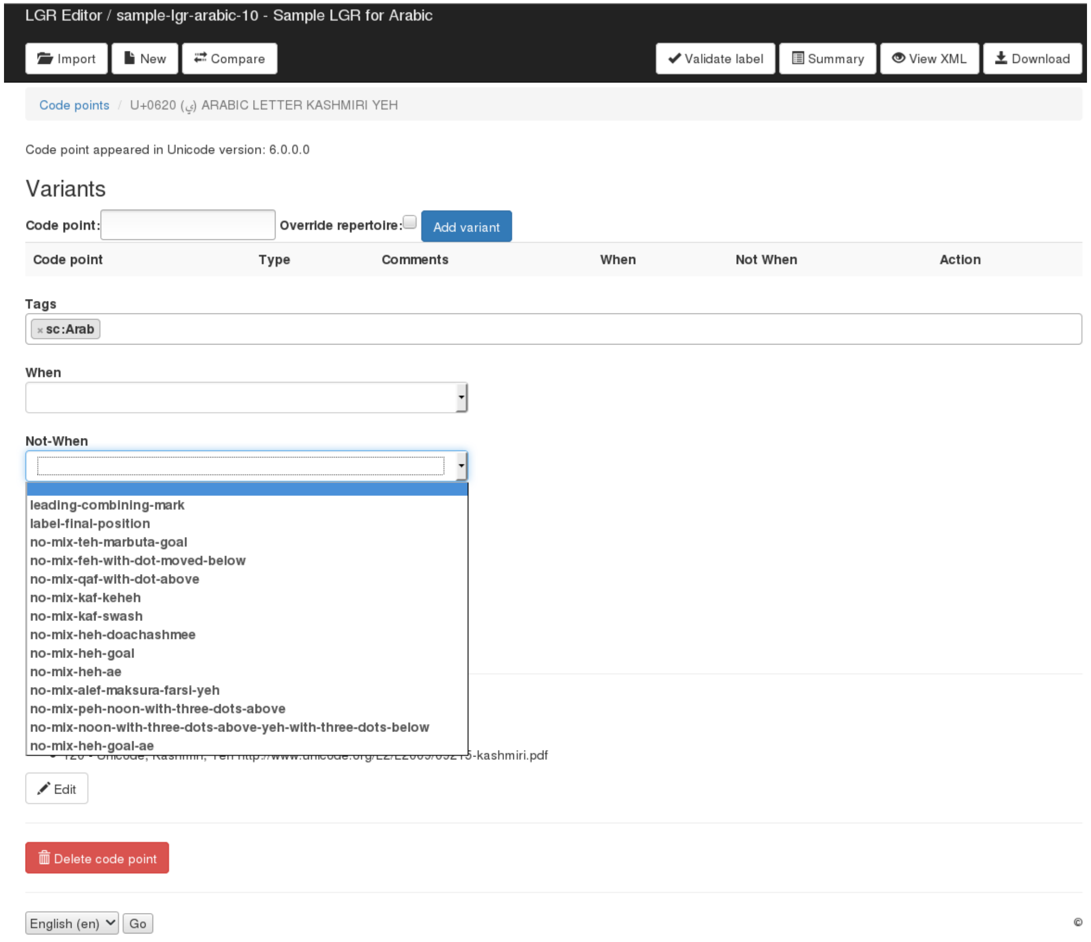


Code point	Character Name	Comments	Action
U+0061 (a) *0 Variant(s)*	LATIN SMALL LETTER A		See code point
U+0062 (b) *0 Variant(s)*	LATIN SMALL LETTER B		See code point
U+0065 (e) … U+006D (m)	LATIN SMALL LETTER E … LATIN SMALL LETTER M		See code point Expand range

Code point	Character Name	Comments	Action
U+0061 (a) *0 Variant(s)*	LATIN SMALL LETTER A		See code point
U+0062 (b) *0 Variant(s)*	LATIN SMALL LETTER B		See code point
U+0065 (e) *0 Variant(s)*	LATIN SMALL LETTER E		See code point
U+0066 (f) *0 Variant(s)*	LATIN SMALL LETTER F		See code point
U+0067 (g) *0 Variant(s)*	LATIN SMALL LETTER G		See code point
U+0068 (h) *0 Variant(s)*	LATIN SMALL LETTER H		See code point
U+0069 (i) *0 Variant(s)*	LATIN SMALL LETTER I		See code point
U+006A (j) *0 Variant(s)*	LATIN SMALL LETTER J		See code point
U+006B (k) *0 Variant(s)*	LATIN SMALL LETTER K		See code point

Fixes Since Phase 2

- Fixes: When/not-when definition in code point, <anchor/> element.



LGR Editor / sample-lgr-arabic-10 - Sample LGR for Arabic

Code points / U+0620 (ؠ) ARABIC LETTER KASHMIRI YEH

Code point appeared in Unicode version: 6.0.0.0

Variants

Code point: Override repertoire: □ Add variant

Code point	Type	Comments	When	Not When	Action

Tags

× sc:Arab

When

Not-When

- leading-combining-mark
- label-final-position
- no-mix-teh-marbuta-goal
- no-mix-feh-with-dot-moved-below
- no-mix-qaf-with-dot-above
- no-mix-kaf-keheh
- no-mix-kaf-swash
- no-mix-heh-doachashmee
- no-mix-heh-goal
- no-mix-heh-ae
- no-mix-alef-maksura-farsi-yeh
- no-mix-peh-noon-with-three-dots-above
- no-mix-noon-with-three-dots-above-yeh-with-three-dots-below
- no-mix-heh-goal-ae

- 120 - Unicode, Kashmiri, Yeh http://www.unicode.org/L2/L2009/09215-kashmiri.pdf

Fixes Since Phase 2

- Security audit performed on code:
  - From report: “The LGR Editor application appeared well-designed, resisting typical web application attacks such as cross-site scripting and URL manipulation. “
  - One issue allowed testing the existence of file (external entities), disabled since then.

Conclusion

Conclusion

- Phase 3 released. LGR manipulation tools.
- Together with bug fixes and some enhancements.
- New work started on collision management.

LGR Considerations

Michel Suignard | IDN LGR Workshop | 9 March 2016

LGR Considerations - Summary

- A well-documented code point repertoire.
- Description of the rules establishing well-formed labels.
  - Correct spelling is not a goal.
- Favor context rules over action-based rules.
- Variants may need to consider related scripts.

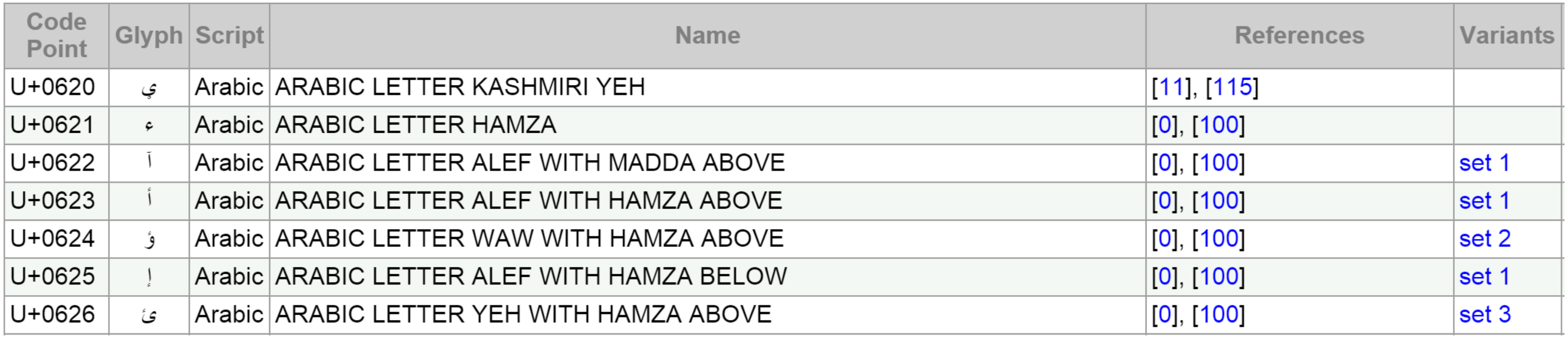
Repertoire

- All code points must:
  - Fall within scope of MSR-2.
  - Have verifiable references.
  - Be referenced the same way in document and XML file.

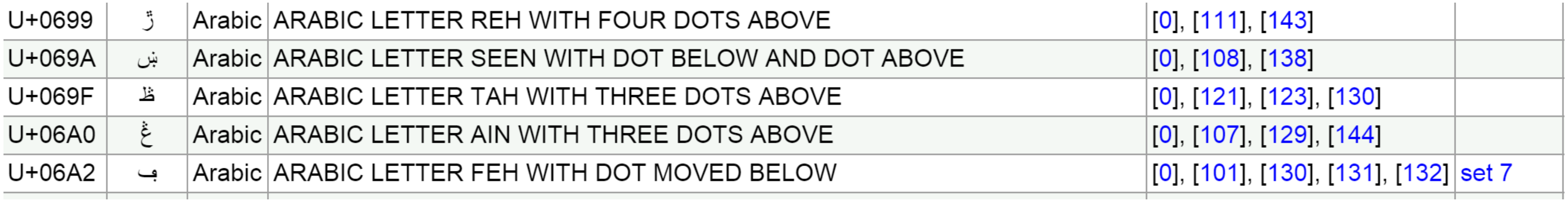
- Repertoire elements can be:
  - Sequences or single code points.
  - Restricted by context (‘when’ and ‘not-when’ rules).
  - Classified (e.g. tag=“vowel’) for use in rules.

- Sequences:
  - Allow specific combination of code points.
  - Are useful for combining marks and other code points that occur only in fixed contexts.
  - Cannot have a tag.

Repertoire



Code Point	Glyph	Script	Name	References	Variants
U+0620	ؠ	Arabic	ARABIC LETTER KASHMIRI YEH	[11], [115]	
U+0621	ء	Arabic	ARABIC LETTER HAMZA	[0], [100]	
U+0622	آ	Arabic	ARABIC LETTER ALEF WITH MADDA ABOVE	[0], [100]	set 1
U+0623	أ	Arabic	ARABIC LETTER ALEF WITH HAMZA ABOVE	[0], [100]	set 1
U+0624	ؤ	Arabic	ARABIC LETTER WAW WITH HAMZA ABOVE	[0], [100]	set 2
U+0625	إ	Arabic	ARABIC LETTER ALEF WITH HAMZA BELOW	[0], [100]	set 1
U+0626	ئ	Arabic	ARABIC LETTER YEH WITH HAMZA ABOVE	[0], [100]	set 3



U+0699	ڙ	Arabic	ARABIC LETTER REH WITH FOUR DOTS ABOVE	[0], [111], [143]	
U+069A	ښ	Arabic	ARABIC LETTER SEEN WITH DOT BELOW AND DOT ABOVE	[0], [108], [138]	
U+069F	ڟ	Arabic	ARABIC LETTER TAH WITH THREE DOTS ABOVE	[0], [121], [123], [130]	
U+06A0	ڠ	Arabic	ARABIC LETTER AIN WITH THREE DOTS ABOVE	[0], [107], [129], [144]	
U+06A2	ڢ	Arabic	ARABIC LETTER FEH WITH DOT MOVED BELOW	[0], [101], [130], [131], [132]	set 7

Well-Formed Labels

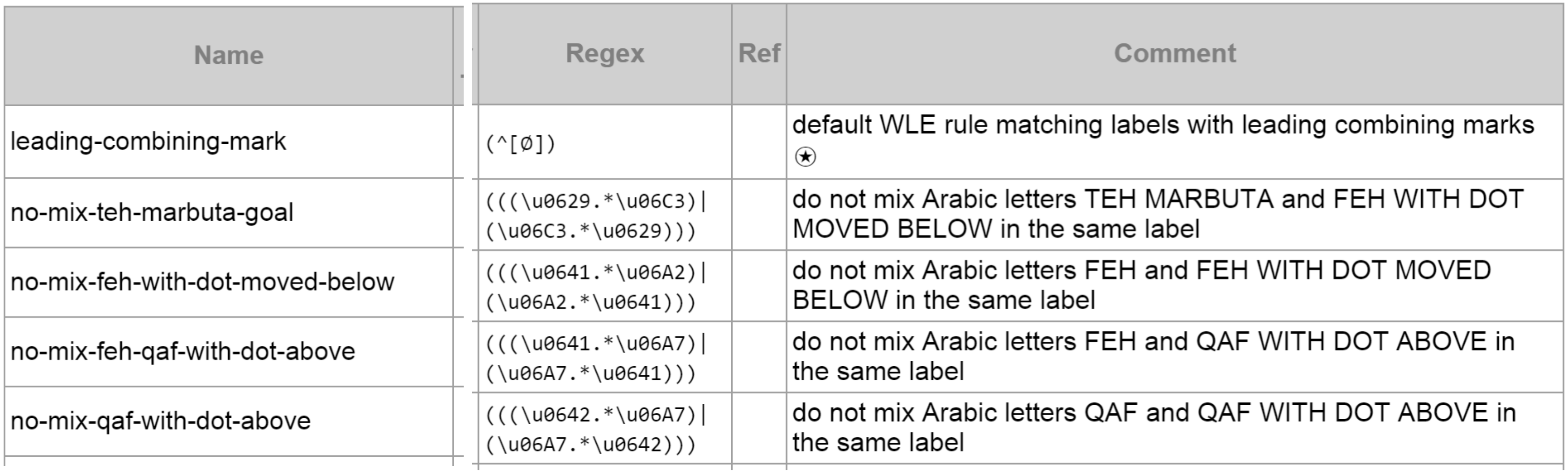
- Specify a set of constraints for syllabic writing systems:
  - Vowels/consonants:
    - Max/min numbers of them.
    - Constraints on order in sequence.
  - Various signs and marks.
- Limit usage of combining marks, when possible.
- Compromise between complexity and coverage.
- Use of BNF or regular expression to express these constraints.
- Use LGRs with similar constraints as templates/models.

Rules

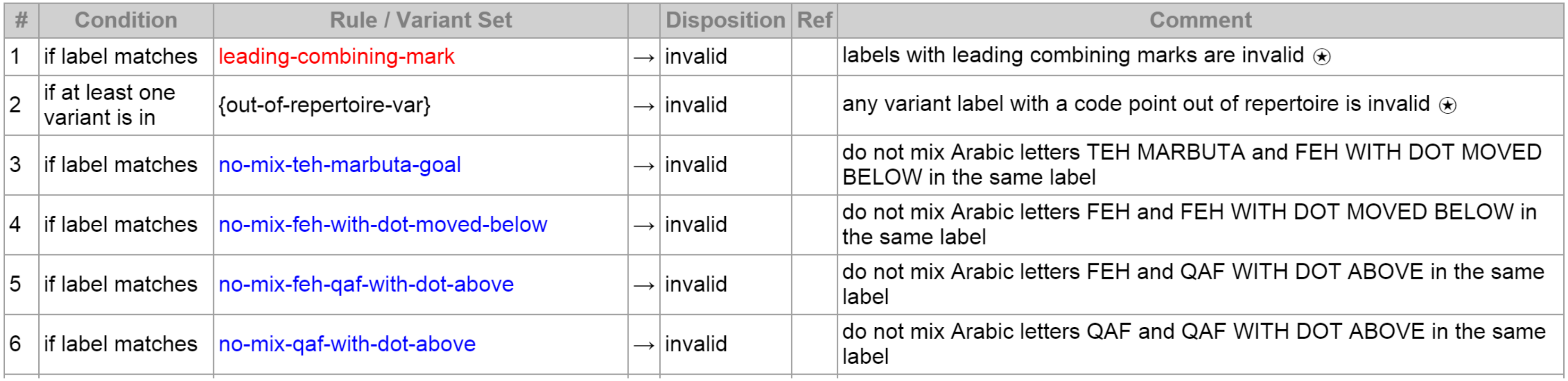
- Enforce well-formed labels.
- Context rules preferred over whole label rules
  - Example: Character must follow a consonant:

    <char cp=“xxxx” when=“follows-consonant” />
- Use tags to classify code points
  - Write rules in terms of sets, not explicit lists of code points.
- Rules can define sequences of code points and be used in another rule.
- Keep it simple.
- Validate with test labels.

Rules



Name	Regex	Ref	Comment
leading-combining-mark	`(^[Ø])`		default WLE rule matching labels with leading combining marks ✪
no-mix-teh-marbuta-goal	`(((\u0629.*\u06C3)|(\u06C3.*\u0629)))`		do not mix Arabic letters TEH MARBUTA and FEH WITH DOT MOVED BELOW in the same label
no-mix-feh-with-dot-moved-below	`(((\u0641.*\u06A2)|(\u06A2.*\u0641)))`		do not mix Arabic letters FEH and FEH WITH DOT MOVED BELOW in the same label
no-mix-feh-qaf-with-dot-above	`(((\u0641.*\u06A7)|(\u06A7.*\u0641)))`		do not mix Arabic letters FEH and QAF WITH DOT ABOVE in the same label
no-mix-qaf-with-dot-above	`(((\u0642.*\u06A7)|(\u06A7.*\u0642)))`		do not mix Arabic letters QAF and QAF WITH DOT ABOVE in the same label



#	Condition	Rule / Variant Set		Disposition	Ref	Comment
1	if label matches	leading-combining-mark	→	invalid		labels with leading combining marks are invalid ✪
2	if at least one variant is in	{out-of-repertoire-var}	→	invalid		any variant label with a code point out of repertoire is invalid ✪
3	if label matches	no-mix-teh-marbuta-goal	→	invalid		do not mix Arabic letters TEH MARBUTA and FEH WITH DOT MOVED BELOW in the same label
4	if label matches	no-mix-feh-with-dot-moved-below	→	invalid		do not mix Arabic letters FEH and FEH WITH DOT MOVED BELOW in the same label
5	if label matches	no-mix-feh-qaf-with-dot-above	→	invalid		do not mix Arabic letters FEH and QAF WITH DOT ABOVE in the same label
6	if label matches	no-mix-qaf-with-dot-above	→	invalid		do not mix Arabic letters QAF and QAF WITH DOT ABOVE in the same label

Test Labels

- Generation Panels have to provide test labels with their LGR proposal.
  - This enables the Integration Panel to test and verify the LGR with appropriate labels.

- Test labels:
  - Content should be comprehensive enough to exercise most cases of the WLE rules, including positive and negative results.
  - Large set is better than small.
  - Use script/language corpus, when available.

Variants

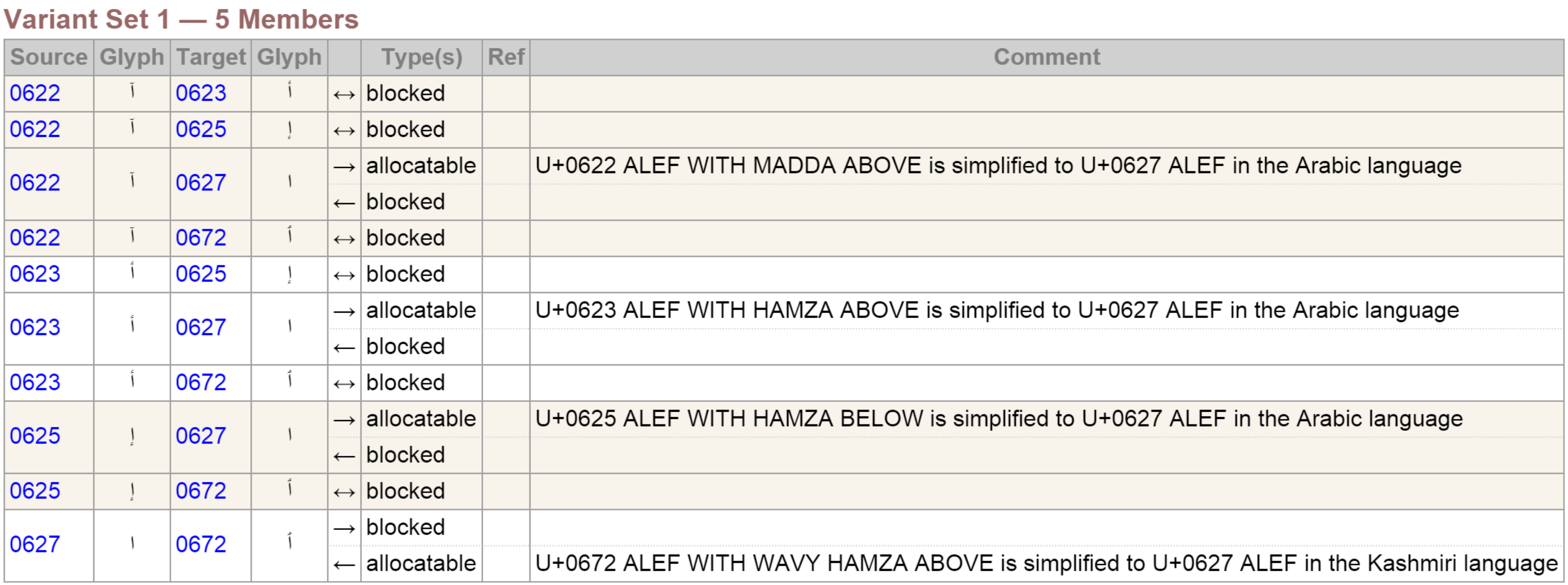
- Variant relation should not depend on accidental similarity
  - Instead, semantic variants or true “homoglyph”.

- Combining marks may need to be considered along with base characters (not in isolation).

- Avoid complicated scenarios
  - Do not use context rules on variants.
  - Other types of confusability will be addressed with different protocols/processes beyond LGRs.

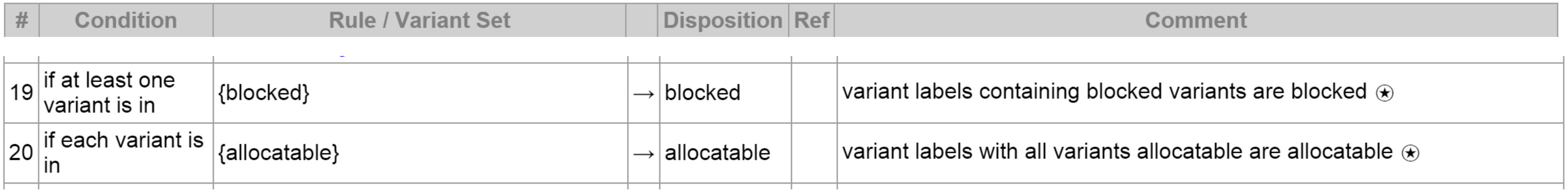
- Inter-script variants needed only when confusion can be foreseen
  - Cases where full labels are indistinguishable across scripts.
  - Such confusion would be between complete labels, not single characters.
  - In practice, different Generation Panels will have to coordinate.

Variants



Variant Set 1 — 5 Members

Source	Glyph	Target	Glyph		Type(s)	Ref	Comment
0622	آ	0623	أ	↔	blocked		
0622	آ	0625	إ	↔	blocked		
0622	آ	0627	ا	→	allocatable		U+0622 ALEF WITH MADDA ABOVE is simplified to U+0627 ALEF in the Arabic language
				←	blocked		
0622	آ	0672	ٲ	↔	blocked		
0623	أ	0625	إ	↔	blocked		
0623	أ	0627	ا	→	allocatable		U+0623 ALEF WITH HAMZA ABOVE is simplified to U+0627 ALEF in the Arabic language
				←	blocked		
0623	أ	0672	ٲ	↔	blocked		
0625	إ	0627	ا	→	allocatable		U+0625 ALEF WITH HAMZA BELOW is simplified to U+0627 ALEF in the Arabic language
				←	blocked		
0625	إ	0672	ٲ	↔	blocked		
0627	ا	0672	ٲ	→	blocked		
				←	allocatable		U+0672 ALEF WITH WAVY HAMZA ABOVE is simplified to U+0627 ALEF in the Kashmiri language



#	Condition	Rule / Variant Set		Disposition	Ref	Comment
19	if at least one variant is in	{blocked}	→	blocked		variant labels containing blocked variants are blocked ✪
20	if each variant is in	{allocatable}	→	allocatable		variant labels with all variants allocatable are allocatable ✪

Example: Arabic Element LGR

- XML file:
  - https://www.icann.org/sites/default/files/lgr/lgr-1-common-01dec15-en.xml

- HTML documentation (extracted from XML file):
  - https://www.icann.org/sites/default/files/lgr/lgr-1-arabic-script-01dec15-en.html

Use Integration Panel Expertise

- Help in formulating the repertoire in terms of code points and references.
- Creation of the repertoire syntax (BNF and regular expression).
- Determination of which constraints on (syllable) structure should be enforced.
- Conversion to XML syntax.
- Validation of the model using test labels and large script corpus.
- Creation of a LGR is better done through successive iteration with feedback from the IP along the way.

Thank You

Resources:

Template for LGR proposals:
https://community.icann.org/download/attachments/43989034/LGR-Proposal-Template.docx

Requirements for LGR Proposals:
https://www.icann.org/en/system/files/files/Requirements-for-LGR-Proposals-20150424.pdf

Packaging the Integrated LGR:
https://community.icann.org/download/attachments/43989034/Packaging-MSR-LGR.pdf

XML Specification for LGRs:
http://www.ietf.org/id/draft-ietf-lager-specification

MSR-2:
https://www.icann.org/en/system/files/files/msr-2-overview-14apr15-en.pdf

ICANN

Thai Script GP Update

Pitinan Kooarmornpatana | IDN LGR Workshop | 9 March 2016

Background: Internet in Thailand

- As of June 30, 2015, according to Internet World Stat: Usage and Population Statistics Report, Thailand has reached 68 million in total population. Only one-third of the total population are active Internet users, since language is critical barrier.

- Thailand has announced the Digital Economy as a road map to enhance its competitive advantage in next five years.

- Therefore, empowering all Thai people to access and use the Internet effectively in order to reduce the digital divide created by the language barrier is needed.

Thai Script

1

ISO 15924

ISO 15924 – Code: Thai
ISO 15924 – Number: 352
ISO 15924 – English name: Thai

2

Unicode Range:

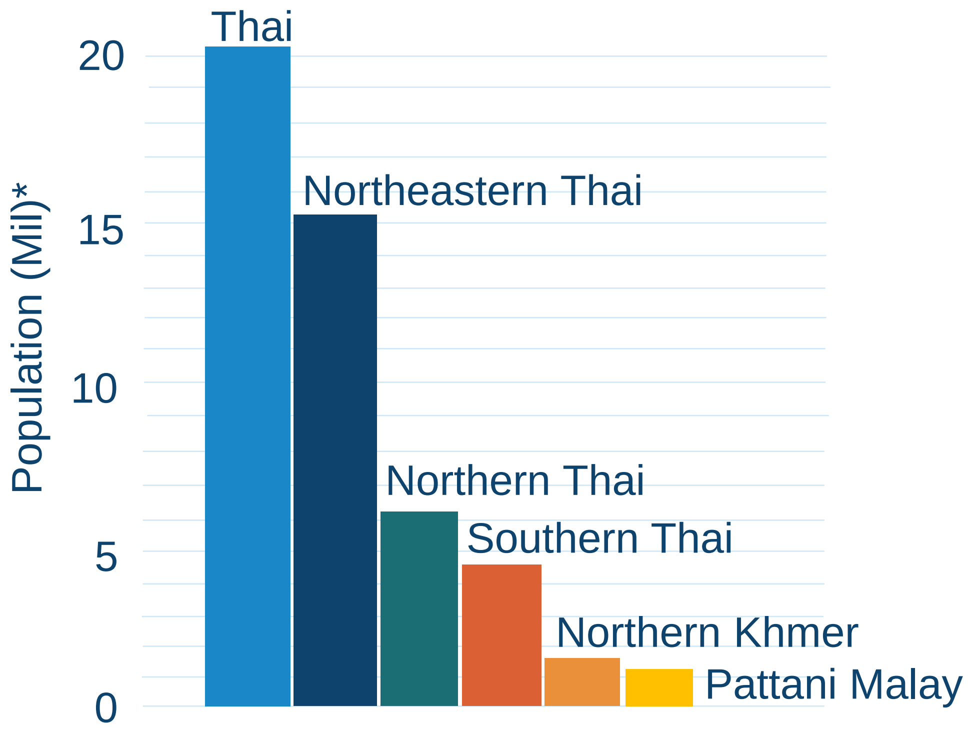
U+0E00 – U+0E7F

3

Writing systems that use Thai script

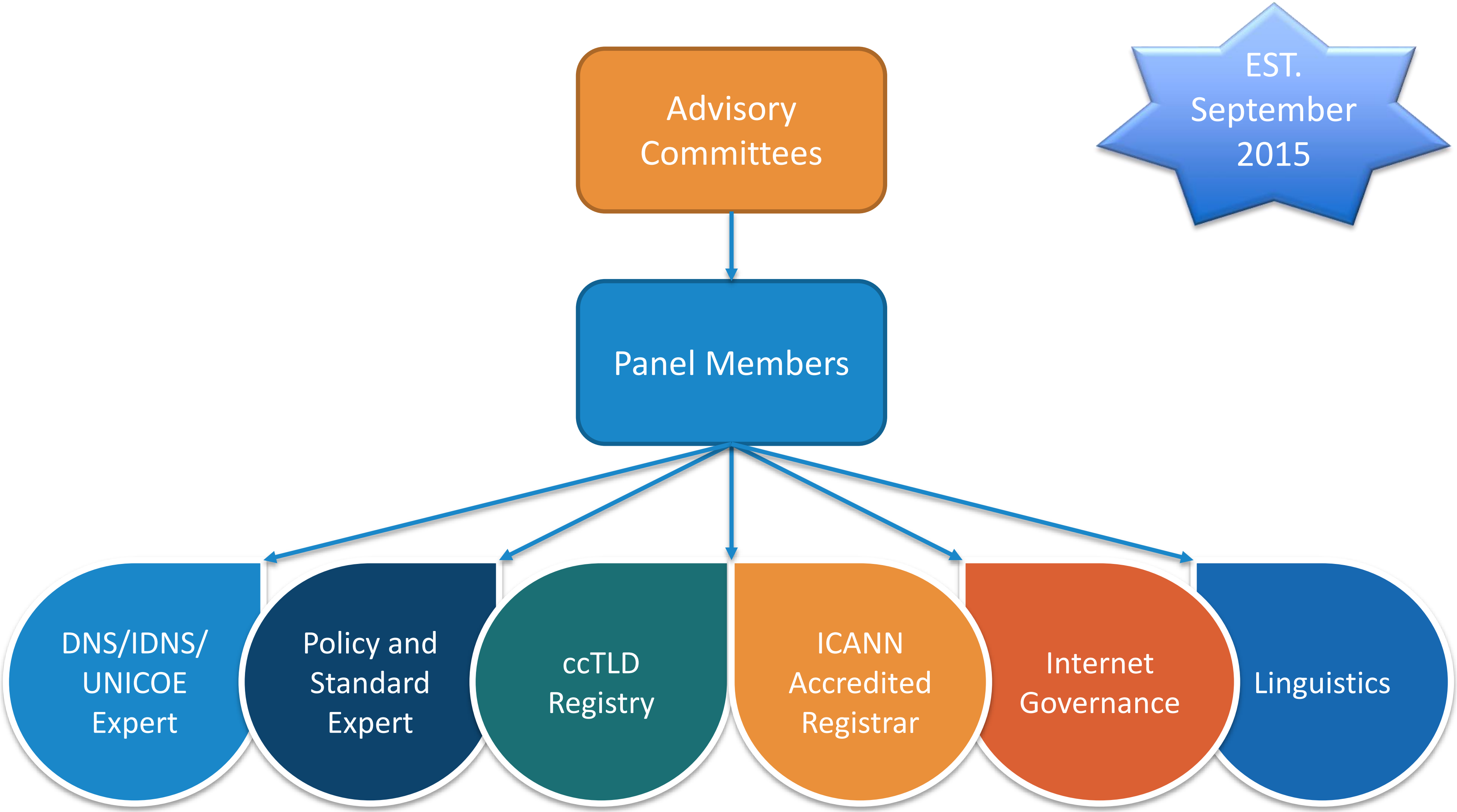
35 languages

Selected Languages Written in **Thai Script**

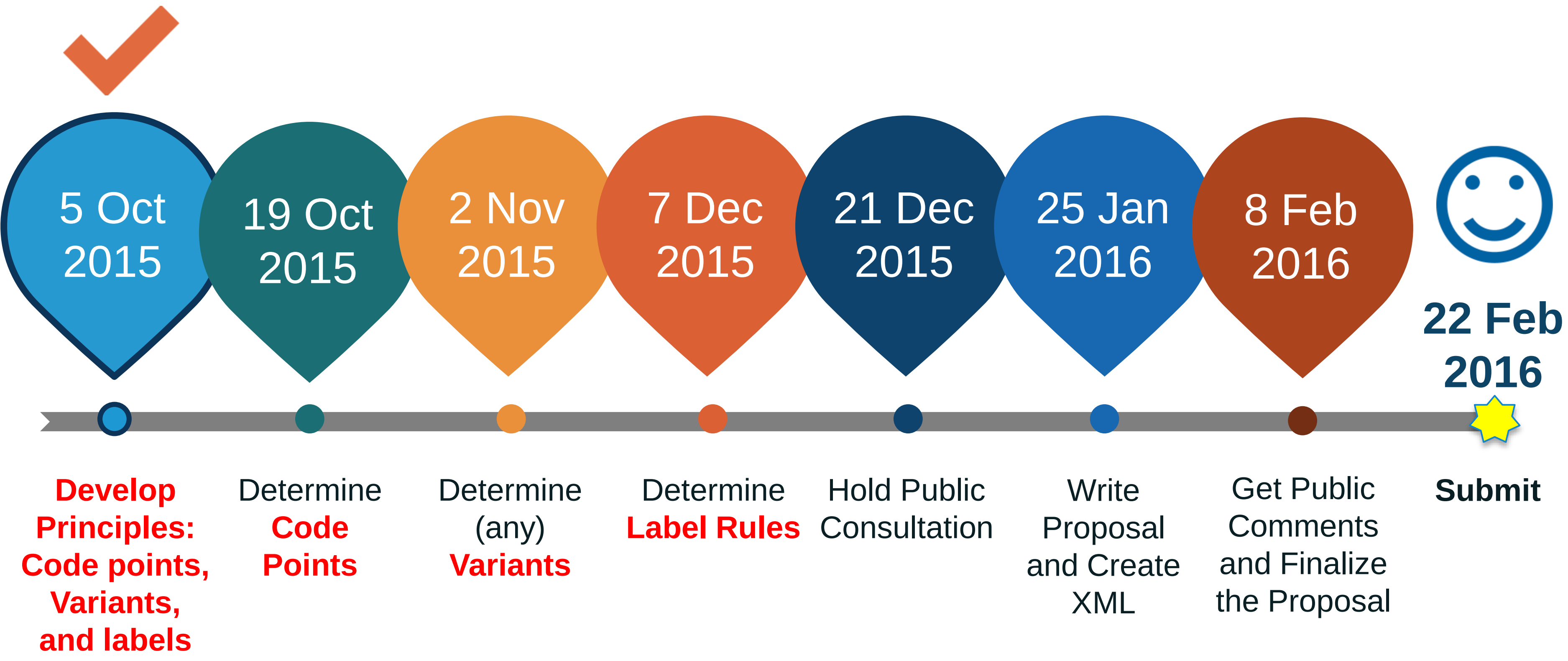


*Source: www.ethnologue.com and scriptsource.org

Thai Generation Panel



Timeline (as of July 2015)



5 Oct 2015

Develop Principles: Code points, Variants, and labels

19 Oct 2015

Determine **Code Points**

2 Nov 2015

Determine (any) **Variants**

7 Dec 2015

Determine **Label Rules**

21 Dec 2015

Hold Public Consultation

25 Jan 2016

Write Proposal and Create XML

8 Feb 2016

Get Public Comments and Finalize the Proposal

22 Feb 2016

Submit

To Summarize

The generation panel will start the work for developing the Root Zone Label Generation Rules (LGR) for Thai scripts by October 2015 and intends to finalize the proposal within February 2016.

Principles for Determining Code Points for Thai Script LGR for the Root Zone

Thai Script

1

ISO 15924

ISO 15924 – Code: Thai
ISO 15924 – Number: 352
ISO 15924 – English name: Thai
Unicode Range: **U+0E00 – U+0E7F**

2

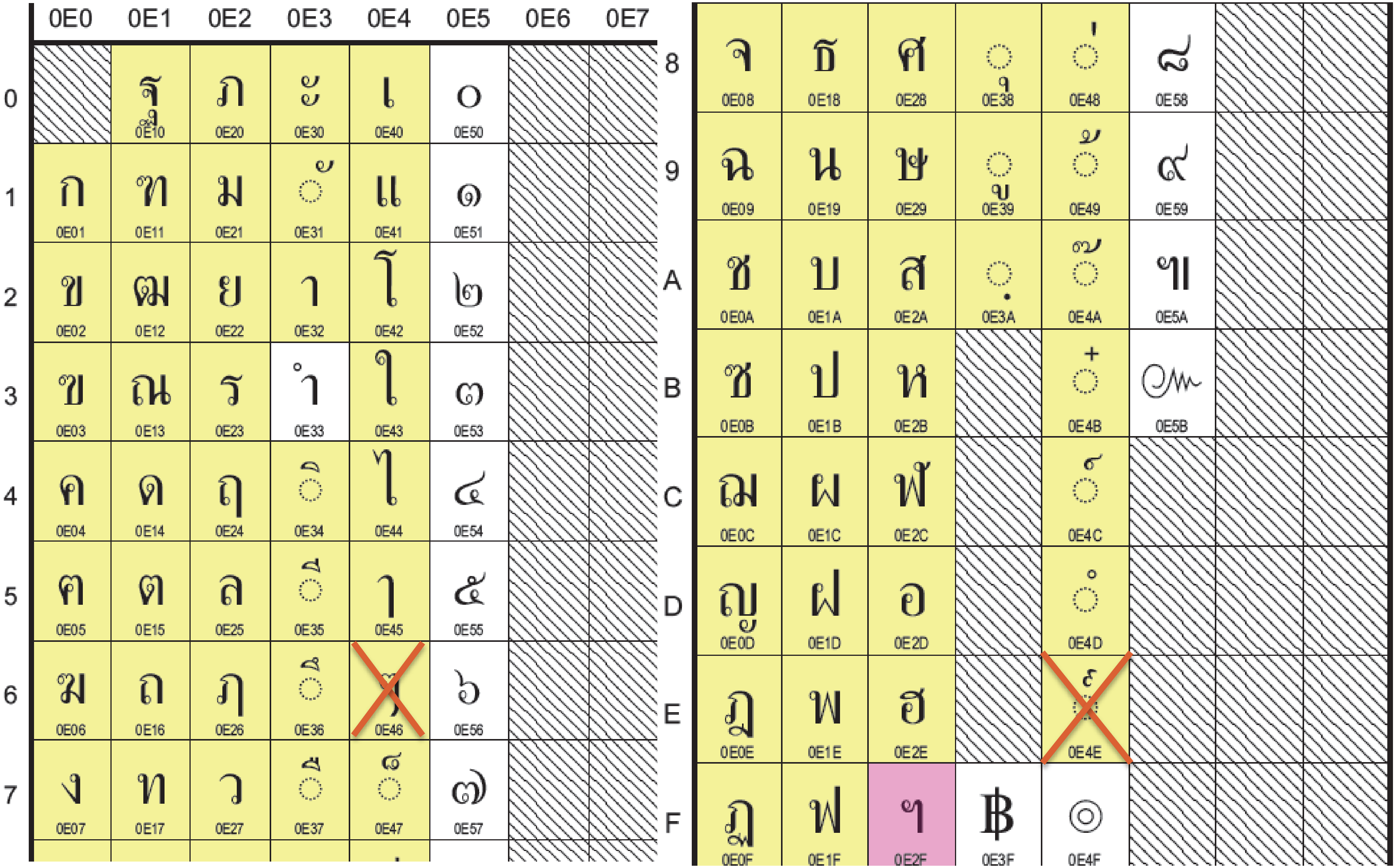
Writing systems that use Thai script

35 languages

Language	ISO 639-3 Code	Locations	Population	Status
Thai	tha	Thailand (official language of Thailand)	20,200,000 (2000)	1
Northeastern Thai	tts	Widespread in Northeast Thailand	15,000,000 (1983 SIL)	6a
Northern Thai	nod	Northern region of Thailand	6,000,000 (1983 SIL)	5
Southern Thai	sou	Southern region of Thailand	4,500,000 (2006 Mahidol University)	5
Northern Khmer	kxm	Northeastern and Eastern regions of Thailand along the border with Cambodia	1,400,000 (2006 Mahidol University)	5
Pattani Malay	mfa	Southern region of Thailand near the border with Malaysia	1,000,000 (2006 Mahidol University)	5

*Source: http://www.ethnologue.com/country/TH/languages
and http://scriptsource.org/cms/scripts/page.php?item_id=script_detail&key=Thai

Finalizing the code point repertoire



	0E0	0E1	0E2	0E3	0E4	0E5	0E6	0E7
0		ฐ 0E10	ภ 0E20	ะ 0E30	เ 0E40	๐ 0E50		
1	ก 0E01	ฑ 0E11	ม 0E21	◌ั 0E31	แ 0E41	๑ 0E51		
2	ข 0E02	ฒ 0E12	ย 0E22	า 0E32	โ 0E42	๒ 0E52		
3	ฃ 0E03	ณ 0E13	ร 0E23	ำ 0E33	ใ 0E43	๓ 0E53		
4	ค 0E04	ด 0E14	ฤ 0E24	◌ิ 0E34	ไ 0E44	๔ 0E54		
5	ฅ 0E05	ต 0E15	ล 0E25	◌ี 0E35	ๅ 0E45	๕ 0E55		
6	ฆ 0E06	ถ 0E16	ฦ 0E26	◌ึ 0E36	~~ๆ~~ ~~0E46~~	๖ 0E56		
7	ง 0E07	ท 0E17	ว 0E27	◌ื 0E37	◌็ 0E47	๗ 0E57		
8	จ 0E08	ธ 0E18	ศ 0E28	◌ุ 0E38	◌่ 0E48	๘ 0E58		
9	ฉ 0E09	น 0E19	ษ 0E29	◌ู 0E39	◌้ 0E49	๙ 0E59		
A	ช 0E0A	บ 0E1A	ส 0E2A	◌ฺ 0E3A	◌๊ 0E4A	๚ 0E5A		
B	ซ 0E0B	ป 0E1B	ห 0E2B		◌๋ 0E4B	๛ 0E5B		
C	ฌ 0E0C	ผ 0E1C	ฬ 0E2C		◌์ 0E4C			
D	ญ 0E0D	ฝ 0E1D	อ 0E2D		◌ํ 0E4D			
E	ฎ 0E0E	พ 0E1E	ฮ 0E2E		~~◌๎~~ ~~0E4E~~			
F	ฏ 0E0F	ฟ 0E1F	ฯ 0E2F	฿ 0E3F	๏ 0E4F			

Principles of Determining Code Point Variants

Defining the Code Point Variant Principles

- ICANN Guidelines
- Proposal for Arabic Script Root Zone LGR (23 August 2015)

Variants within Repertoire

Two code points are variants if they are visually same as each other

VS

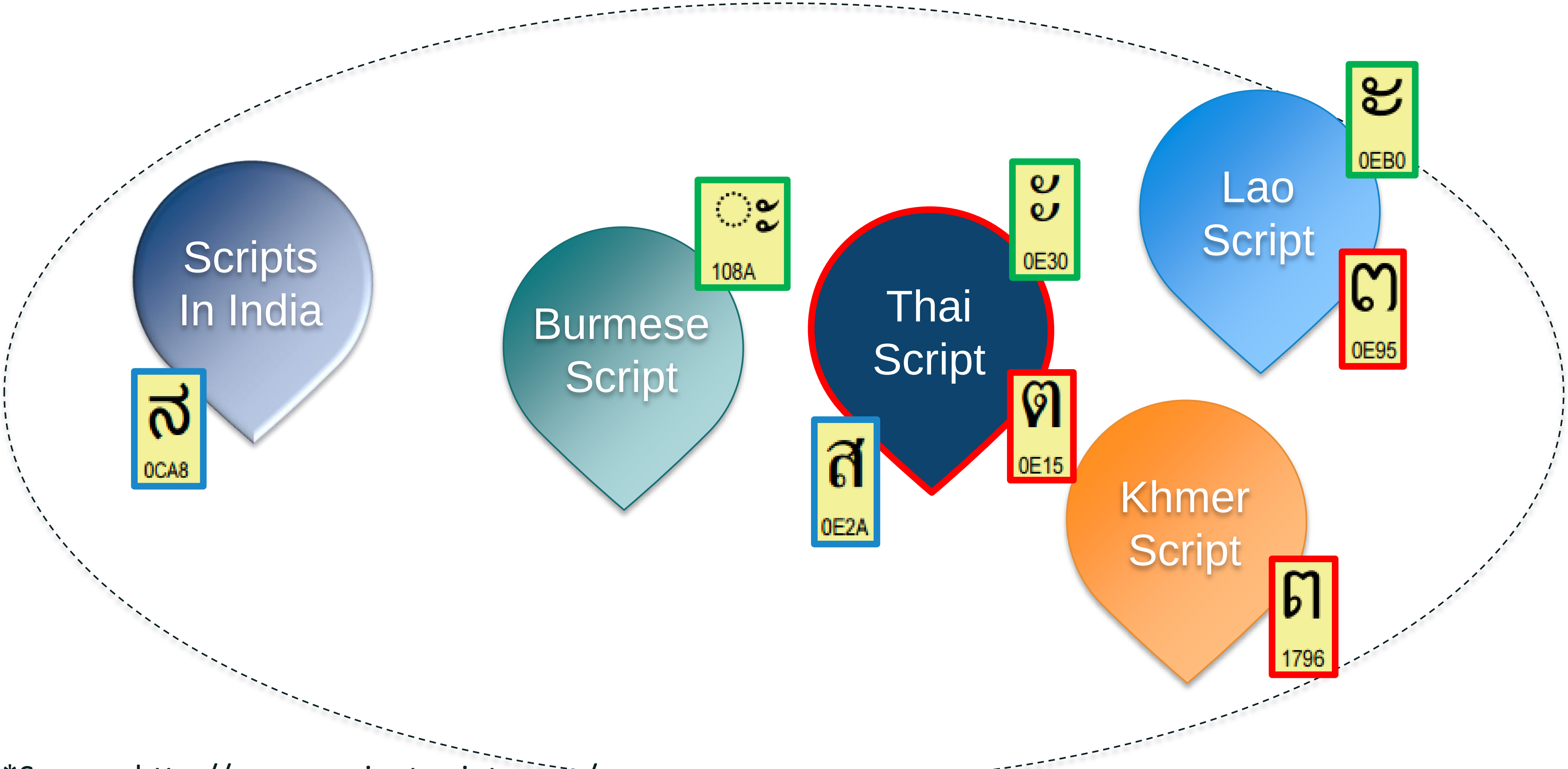
Handling Out-of-Repertoire Variants

At first it may seem counterintuitive to define variants that map to code points not part of the repertoire.

However, for zones for which multiple LGRs are defined, there may be situations where labels valid under one LGR should be blocked if a label under another LGR is already delegated.

Thai Script and Related Scripts

Brahmi script
Khmer script → **Thai script***



*Source: http://www.ancientscripts.com/

Variant within Thai Script

Item #	Unicode Code Point	Glyph	Name and GC	Variants	
				Thai	Unicode
2	0E02	ข	THAI CHARACTER; KHO KHAI	ฃ ช ซ	0E03 0E0A 0E0B
3	0E03	ฃ	THAI CHARACTER; KHO KHUAT	ข ช ซ	0E02 0E0A 0E0B
4	0E04	ค	THAI CHARACTER; KHO KHWAI	ฅ ด ต ศ	0E05 0E14 0E15 0E28
5	0E05	ฅ	THAI CHARACTER; KHO KHON	ค ด ต ศ	0E04 0E14 0E15 0E28
6	0E06	ฆ	THAI CHARACTER; KHO RAKHANG	ม	0E21
10	0E0A	ช	THAI CHARACTER; CHO CHANG	ข ฃ ซ	0E02 0E03 0E0B
11	0E0B	ซ	THAI CHARACTER; SO SO	ข ฃ ช	0E02 0E03 0E0A
12	0E0C	ฌ	THAI CHARACTER; CHO CHOE	ณ	0E13
14	0E0E	ฎ	THAI CHARACTER; DO CHADA	ฏ	0E0F
15	0E0F	ฏ	THAI CHARACTER; TO PATAK	ฎ	0E0E
17	0E11	ฑ	THAI CHARACTER; THO NANGMONTHO	ท	0E17
19	0E13	ณ	THAI CHARACTER; NO NEN	ฌ	0E0C
20	0E14	ด	THAI CHARACTER; DO DEK	ค ฅ ต ศ	0E04 0E05 0E15 0E28
21	0E15	ต	THAI CHARACTER; TO TAO	ค ฅ ด ศ	0E04 0E05 0E14 0E28
22	0E16	ถ	THAI CHARACTER; THO THUNG	ฤ	0E24
23	0E17	ท	THAI CHARACTER; THO THAHAN	ฑ	0E11
26	0E1A	บ	THAI CHARACTER; BO BAIMAI	ป ษ	0E1B 0E29
27	0E1B	ป	THAI CHARACTER; PO PLA	บ ษ	0E1A 0E29
28	0E1C	ผ	THAI CHARACTER; PHO PHUNG	ฝ	0E1D

29	0E1D	ฝ	THAI CHARACTER; FO FA	ผ	0E1C
30	0E1E	พ	THAI CHARACTER; PHO PHAN	ฟ ฬ	0E1F 0E2C
31	0E1F	ฟ	THAI CHARACTER; FO FAN	พ ฬ	0E1E 0E2C
33	0E21	ม	THAI CHARACTER; MO MA	ฆ	0E06
36	0E24	ฤ	THAI CHARACTER; RU	ถ	0E16
37	0E25	ล	THAI CHARACTER; LO LING	ส	0E2A
40	0E28	ศ	THAI CHARACTER; SO SALA	ค ต ฅ ด	0E04 0E05 0E14 0E15
41	0E29	ษ	THAI CHARACTER; SO RUSI	บ ป	0E1A 0E1B
42	0E2A	ส	THAI CHARACTER; SO SUA	ล	0E25
44	0E2C	ฬ	THAI CHARACTER; LO CHULA	พ ฟ	0E1E 0E1F
45	0E2D	อ	THAI CHARACTER; O ANG	ฮ	0E2E
46	0E2E	ฮ	THAI CHARACTER; HO NOKHUK	อ	0E2D
49	0E32	า	THAI CHARACTER; SARA AA	ๅ	0E45
50	0E33	◌ำ	THAI CHARACTER; SARA AM	◌ํ + า	0E4D + 0E32
51	0E34	◌ิ	THAI CHARACTER; SARA I	◌ี ◌ึ ◌ื	0E35 0E36 0E37
52	0E35	◌ี	THAI CHARACTER; SARA II	◌ิ ◌ึ ◌ื	0E34 0E36 0E37
53	0E36	◌ึ	THAI CHARACTER; SARA UE	◌ิ ◌ี ◌ื	0E34 0E35 0E37
54	0E37	◌ื	THAI CHARACTER; SARA UEE	◌ิ ◌ี ◌ึ	0E34 0E35 0E36
55	0E38	◌ุ	THAI CHARACTER; SARA U	◌ู	0E39
56	0E39	◌ู	THAI CHARACTER; SARA UU	◌ุ	0E38
59	0E41	แ	THAI CHARACTER; SARA AE	เ + เ	0E40 + 0E40
63	0E45	ๅ	THAI CHARACTER; LAKKHANGYAO	า	0E32

Variant within Laos Script

Lao Character	Unicode	Thai Character	Unicode
ກ	0E81	ท	0E17
ຄ	0E84	ถ	0E16
ຈ	0E88	จ	0E08
ຍ	0E8D	ย	0E22
ດ	0E94	ถ	0E16
ຕ	0E95	ต	0E15
ຖ	0E96	ถ ฤ	0E16 0E24
ທ	0E97	ท	0E17
ນ	0E99	ม	0E21
ບ	0E9A	บ	0E1A
ປ	0E9B	ป	0E1B
ຜ	0E9C	ผ	0E1C
ຝ	0E9D	ฝ	0E1D
ພ	0E9E	พ	0E1E
ຟ	0E9F	ฟ	0E1F
ມ	0EA1	ม	0E21
ຢ	0EA2	ย	0E22
ຣ	0EA3	ธ ร	0E18 0E23
ລ	0EA5	ล	0E25
ວ	0EA7	ว อ	0E27 0E2D
ສ	0EAA	ส	0E2A
ຫ	0EAB	ท ห	0E17 0E2B
ອ	0EAD	ฮ	0E2E
ຮ	0EAE	ธ ร	0E18 0E23
ະ	0EB0	ะ	0E30
◌ັ	0EB1	◌ั	0E31
າ	0EB2	า	0E32
◌ຳ	0EB3	◌ำ	0E33
◌ື	0EB7	◌๊	0E4A
◌ຸ	0EB8	◌ุ	0E38
◌ູ	0EB9	◌ู	0E39
◌ົ	0EBB	◌์	0E4C
ເ	0EC0	เ	0E40
ແ	0EC1	แ	0E41
ໂ	0EC2	โ	0E42
ໃ	0EC3	ใ	0E43
ໄ	0EC4	ไ	0E44
◌່	0EC8	◌่	0E48
◌້	0EC9	◌้	0E49
◌໊	0ECA	◌๊	0E4A
◌໋	0ECB	◌๋	0E4B
◌໌	0ECC	◌์	0E4C
◌ໍ	0ECD	◌ํ	0E4D
ໟ	0EDF	ย ษ	0E22 0E29

Variant with Khmer script

Khmer script	Unicode	Thai Character	Unicode
ក	1780	ก + ◌็	0E01 + 0E47
គ	1782	ค+◌็	0E04+0E47
ឃ	1783	ช+ช ฬ	0E0A+0E0A 0E2C
ង	1784	ฬ	0E2C
ឈ	1788	ถ+ช+ช	0E16+0E0A+0E0A
ដ	178A	ผ+◌ั	0E1C+0E31
ឋ	178B	ช ซ	0E0A 0E0B
ឍ	178D	ฌ ฒ ต+ร	0E0C 0E12 0E15+0E23
ណ	178E	ฌ+ก ญ	0E0C+0E01 0E0C
ត	178F	ด+◌๊	0E14+0E4A
ន	1793	ร	0E23
ប	1794	ช ย	0E0A 0E22
ព	1796	ด ต ถ	0E14 0E15 0E16
ភ	1797	ภ + ◌็	0E20 + 0E47
ម	1798	ษ ย	0E29 0E22
យ	1799	ผ	0E1C
រ	179A	ร	0E23
ល	179B	ญ	0E0C
វ	179C	ร	0E23
ឝ	179D	ศ + ◌็	0E28+0E47
ឞ	179E	ษ ย	0E29 0E22
ហ	17A0	ย+า	0E22+0E32
អ	17A2	ร+ร	0E23+0E23
ឥ	17A5	ส	0E2A
ឫ	17AB	ช+◌ุ ย+◌ุ	0E0A+0E38 0E22+0E38
ឬ	17AC	ช+◌ุ	0E0A+0E38
ឭ	17AD	ฅ+◌ุ ต+◌ุ	0E05+0E38 0E15+0E38
ឮ	17AE	ฅ+◌ุ ต+◌ุ	0E05+0E38 0E15+0E38
ឯ	17AF	ฉ ฬ	0E09 0E2C
◌ា	17B6	า ๅ	0E32 0E45
◌ិ	17B7	◌ิ	0E34
◌ី	17B8	◌ี	0E35
◌ឹ	17B9	◌ึ	0E36
◌ឺ	17BA	◌ื	0E37
◌ុ	17BB	◌ุ	0E38
◌ូ	17BC	◌ู	0E39
◌ួ	17BD	◌ู	0E39
◌ើ	17BE	เ+◌ี	0E40+0E35
◌េ	17C1	เ	0E40
◌ោ	17C4	เ+า เ+ๅ	0E40+0E32 0E40+0E45
◌ៅ	17C5	เ+า เ+ๅ	0E40+0E32 0E40+0E45
◌ំ	17C6	◌ํ	0E4D
◌ះ	17C7	ะ	0E30
◌ៈ	17C8	ะ	0E30
◌៊	17CA	◌๊	0E4A
◌់	17CB	◌่	0E48
◌៌	17CC	◌็	0E47
◌៍	17CD	◌์	0E4C
◌៎	17CE	◌๋	0E4B
◌៏	17CF	◌็	0E47
◌័	17D0	◌ั	0E31

Variant within Myanmar Script

Myanmar script	Unicode	Thai Character	Unicode
▯	1001	ว อ	0E27 0E2D
▯	1002	ก	0E01
▯	1003	พ ฟ	0E1E 0E1F
▯	1008	ข ฃ	0E02 0E03
▯	100E	ข ฃ บ	0E02 0E03 0E1A
▯	1015	ข บ	0E02 0E1A
▯	1018	ว + ว	0E27 + 0E27
▯	101A	ผ พ	0E1C 0E1E
▯	101B	ฤ ใ	0E24 0E43
▯	102B	า ๅ	0E32 0E45
▯	102D	◌ํ	0E4D
▯	1036	◌ํ	0E4D
▯	1037	ฺ	0E3A
▯	1038	ะ	0E30
▯	1062	า	0E32
▯	1064	า	0E32
▯	1075	ภ	0E20
▯	1076	ว	0E27
▯	1077	ภ	0E20
▯	1080	ม	0E21
▯	108A	ะ	0E30

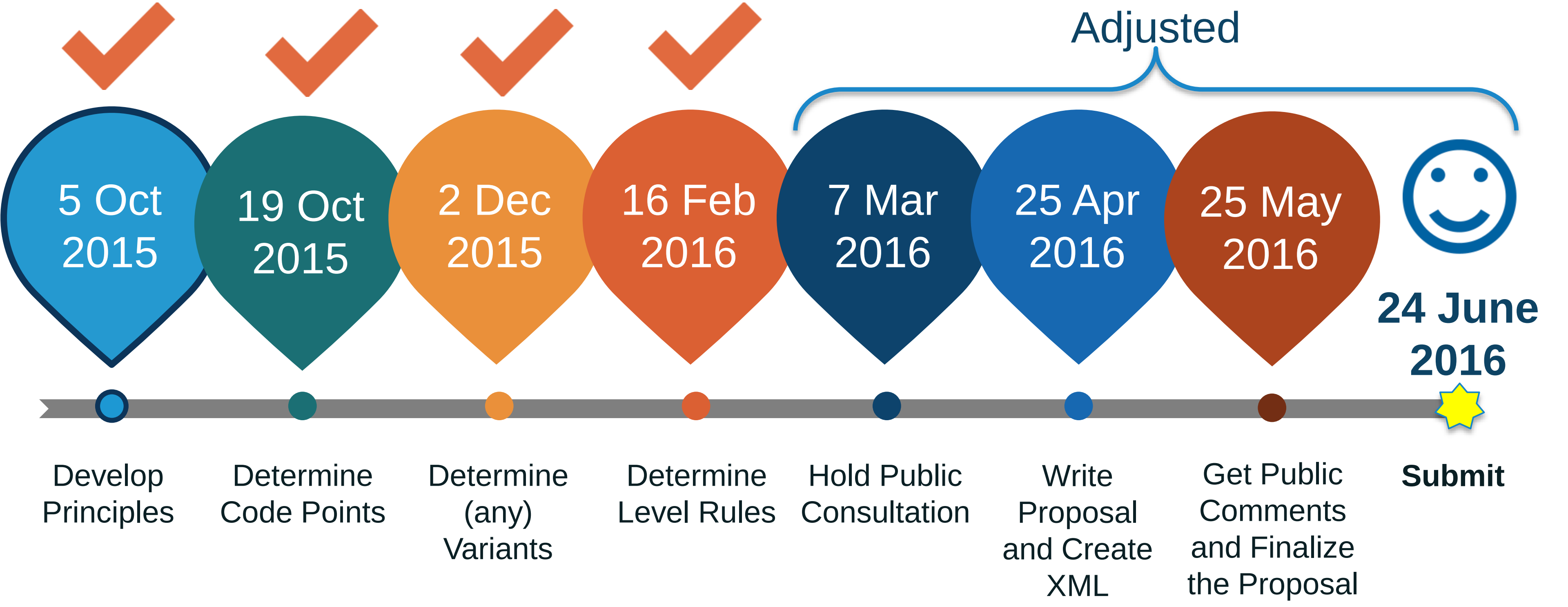
Principles of Determining Whole Label Evaluation Rule

WLE Rule

- Are there sequences of code points that are only valid in a certain order or fixed sequences?
- Can certain code points only appear in a certain position within a label?
- Should certain code points be prevented from appearing in a certain position in a label?
- What is the complexity cost of including a rule?
  - Do related scripts share the same (or a similar) rule?
- What is the risk of not having such a rule?
  - What is the risk of having a simplified / less complex version of the rule?
- Would any defined variants have a different disposition depending on context?
- Are any rules in tension with any of the principles?

Roles and Responsibilities for Task Division

Timeline Thai Script LGR



Adjusted

5 Oct 2015 — Develop Principles

19 Oct 2015 — Determine Code Points

2 Dec 2015 — Determine (any) Variants

16 Feb 2016 — Determine Level Rules

7 Mar 2016 — Hold Public Consultation

25 Apr 2016 — Write Proposal and Create XML

25 May 2016 — Get Public Comments and Finalize the Proposal

24 June 2016 — **Submit**

To Summarize

The generation panel will start the work for developing the Root Zone Label Generation Rules (LGR) for Thai scripts by October 2015 and intends to finalize the proposal within June 2016.

Korean GP Update

KIM Kyongsok | IDN LGR Workshop | 9 March 2016

Agenda

- Introduction and a list of Hangul Syllables for K-LGR v0.3.
- A list of Hangul Syllables, Hanja characters for K-LGR v0.3.
- Review of C (Chinese) and K (Korean) Variant Groups.
- Timeline of KLGP activities.

1. Introduction

- Characters to be included in "kore" (Korean Label)
  - Both Hangeul (Hangul) and Hanja are included.
- K-LGR v0.3 (2015.08.13.)

2. K-LGR v0.3

- **A list of Hangul Syllables for K-LGR v0.3 (2015.08.13.)**
  - 11,172 Hangul Syllbles (U+AC00 ~ U+D7A3)

- **A list of Hanja characters for K-LGR v0.3 (2015.08.13.)**

Source of Hanja Character Set	**# chars**
1) KS X 1001 (268 comptb. chars excluded)	4,620
2) KPS 9566	4,653
3) IICORE - K column marked	4,743
4) IICORE - KP column marked (= KPS 9566)	4,653
5) Qualifying Test of Korean Hanja Proficiency (한국 한자 능력 검정 시험)	4,641
K-LGR v0.3 (2015.08.13.): Hanja List	4,819

3. Review of C (Chinese) and K (Korean) Variant Groups

- C-LGR (2015.04.30.): 3093 variant groups
  (a variant group is composed of two or more variants)
- K-LGR v0.3 (2015.08.13.): 37 variant groups

- **Analysis of 3093 C (Chinese) variant groups**
  - Extracted 303 variant groups where there are two or more K characters
    - K character is a character belonging to K-LGR v0.3 (2015.08.13.)
  - Korea classified 303 variant groups into three categories

3. Review of C (Chinese) and K (Korean) Variant Groups (2)

K position	# variant groups
Acceptable	44
Unacceptable	259
Total	303

- 1) K characters in some C variant groups have different meanings in Korea.

- 2) K characters in some C variant groups have similar meanings; however, those K chars are not regarded as “variants in the context of TLD” in Korea.

3. Review of C (Chinese) and K (Korean) Variant Groups (3)

K chars in 259 Unacceptable C var. group have	# C variant groups
Similar meaning	97
Different meanings	162
Total	259

- Need to translate meanings of K chars with different meanings in 162 unacceptable C variant groups.

3. Review of C (Chinese) and K (Korean) Variant Groups (4)

- **A special class of variant groups in C-LGR**
  - About 56 "Simplitional chars”: [= SIMPLIfied + tradiTIONAL]
    - Currently, the char is a simplified char in China.
    - However, the char had been used for a long time before PRC announced simp. chars in 1964 in Korea, China, etc.
  - An example of Simplitional char: 机

    1) In China:
    - 机: Currently, Simplified char, "machine".
    - 机: simplified from Traditional char 機 (machine).

    2) In Korea: the two chars are distinct
    - 机: desk (reading "gwe")
    - 機: machine (reading "gi")

3. Review of C (Chinese) and K (Korean) Variant Groups (5)

- **Analysis of 37 K (Korean) variant groups and Related 37 C variant groups**
  - In all 37 K variants groups composed of two characters, there are two C characters
    - C character is a character belonging to C-LGR.
  - Korea classified 37 Related C variant groups into three categories as shown below.

K position RE: related C variant groups	# of Related C variant groups
Acceptable	33
Unacceptable *	3
Need to review	1
Total	37

3. Review of C (Chinese) and K (Korean) Variant Groups (6)

○ E.g.1, Related C variant group

- (O 4EC7 仇) (O 8B8E 讎) (O 8B90 讐) (X 96E0 雠)

--> K position: There is much difference in meaning in K between 4EC7 and (8B8E = 8B90).

○ E.g., 2: Related C variant group

- (O 88CF 裏) (O 88E1 裡) (O 91CC 里)

--> C included 91CC since it is a simplified char of traditional characters 88CF and 88E1.

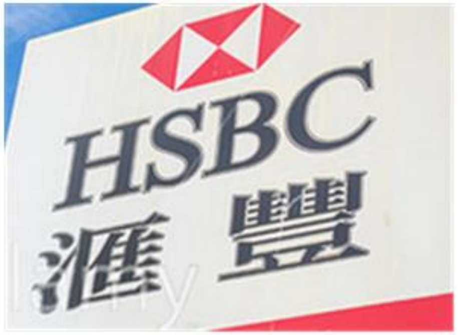
--> K position: There is much difference in meaning in K between 91CC and (88CF = 88E1).

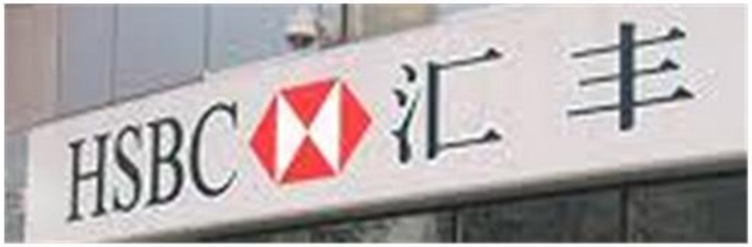
3. Review of C (Chinese) and K (Korean) Variant Groups (7)

- **Possible errors in C-LGR-1 (2015.04.30.)**
  - KLGP reviewed C-LGR-1 (2015.04.30.) and found possible errors.
  - Sent to C members on July 16, 2015 and a few more times later. In two cases, there are symmetry and/or transitivity issues (problems).
    - 矿(77FF) 礦(7926) 砿(783F) 鉱(9271) 鑛(945B)
    - 铁(94C1) 鐵(9435) 鉄(9244) 銕(9295) 鐡(9421)
    - Hope that these have been fixed by now.

3. Review of C (Chinese) and K (Korean) Variant Groups (8)

- **A mixture of trad. and simp. chars.**





  - CGP will allow a HSBC domain composed of mixed simplified and traditional chars @ ICANN meeting in Buenos Aires.
  - left figure: only traditional chars.
  - right figure: only simplified chars.
  - a figure of HSBC domain composed of mixed simplified and traditional chars?

4. KGP’s Activities History (1)

2013

Dec : organization of Korean LGP

2014

Mar : Participate CJK joint meeting @ ICANN49 Singapore

Jun : Participate ICANN50 London and KGP status update

Jul : 1st KGP meeting

Aug : 2nd KGP meeting

Oct : Participate ICANN51 LA and KGP status update

2015

Jan : 3rd KGP meeting and re-composition KGP

Feb : Participate ICANN52 Singapore and KGP status update

Apr : 4th , 5th KGP meeting (reorganization of KGP)

May : 6th , 7th KGP meeting(K-LGR-1 v0.1) and CJK Joint meeting in Seoul

Jun : 8th KGP meeting(K-LGR-1 v0.2) and participate ICANN53 BA

Jul : 9th KGP meeting, workshop and Participate APrIGF Macau

Aug : 10th KGP meeting(K-LGR-1 v0.3)

Sep : 11th KGP meeting

Oct : Call for formal Generation Panel to ICANN and participate ICANN54 Dublin

4. KGP's Activities History (2)

2015 Nov: 12th KGP meeting

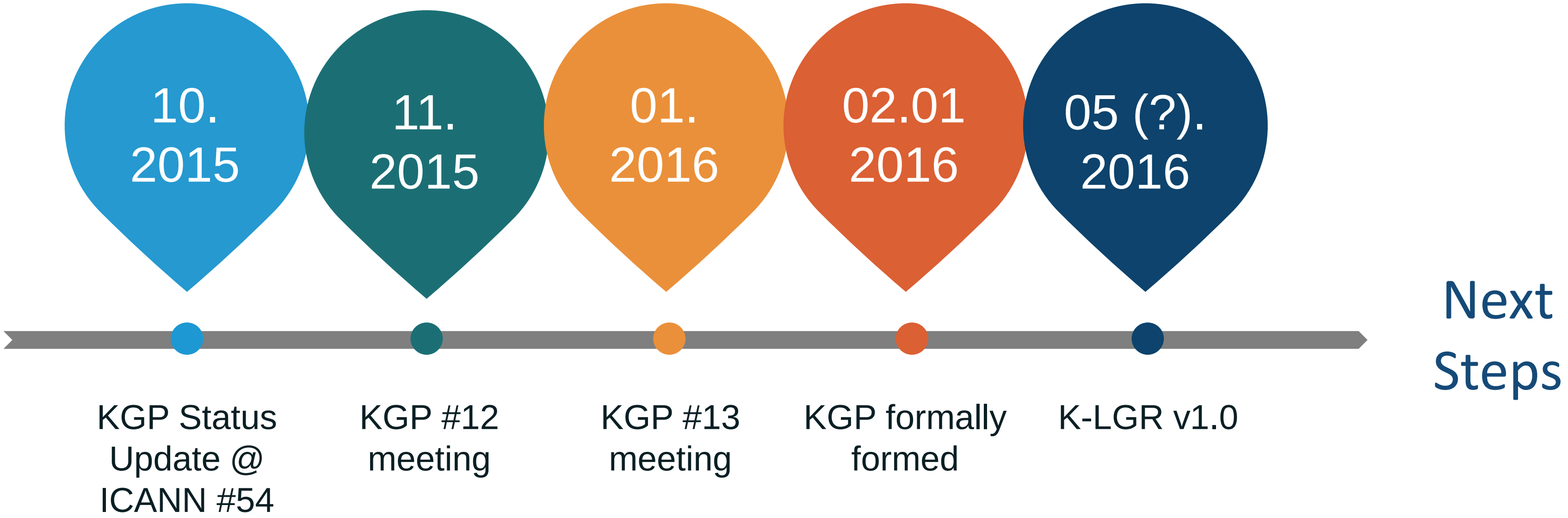
2016

Jan : 13th KGP meeting

Feb : The Korean Community "formally" Forms Generation Panel for Developing the Root Zone Label Generation Rules (LGR), 2016-02-01.

Mar : Participate ICANN55 Marrakesh, Morocco and present KGP status update

5. Timeline of KLGP activities



10. 2015 — KGP Status Update @ ICANN #54

11. 2015 — KGP #12 meeting

01. 2016 — KGP #13 meeting

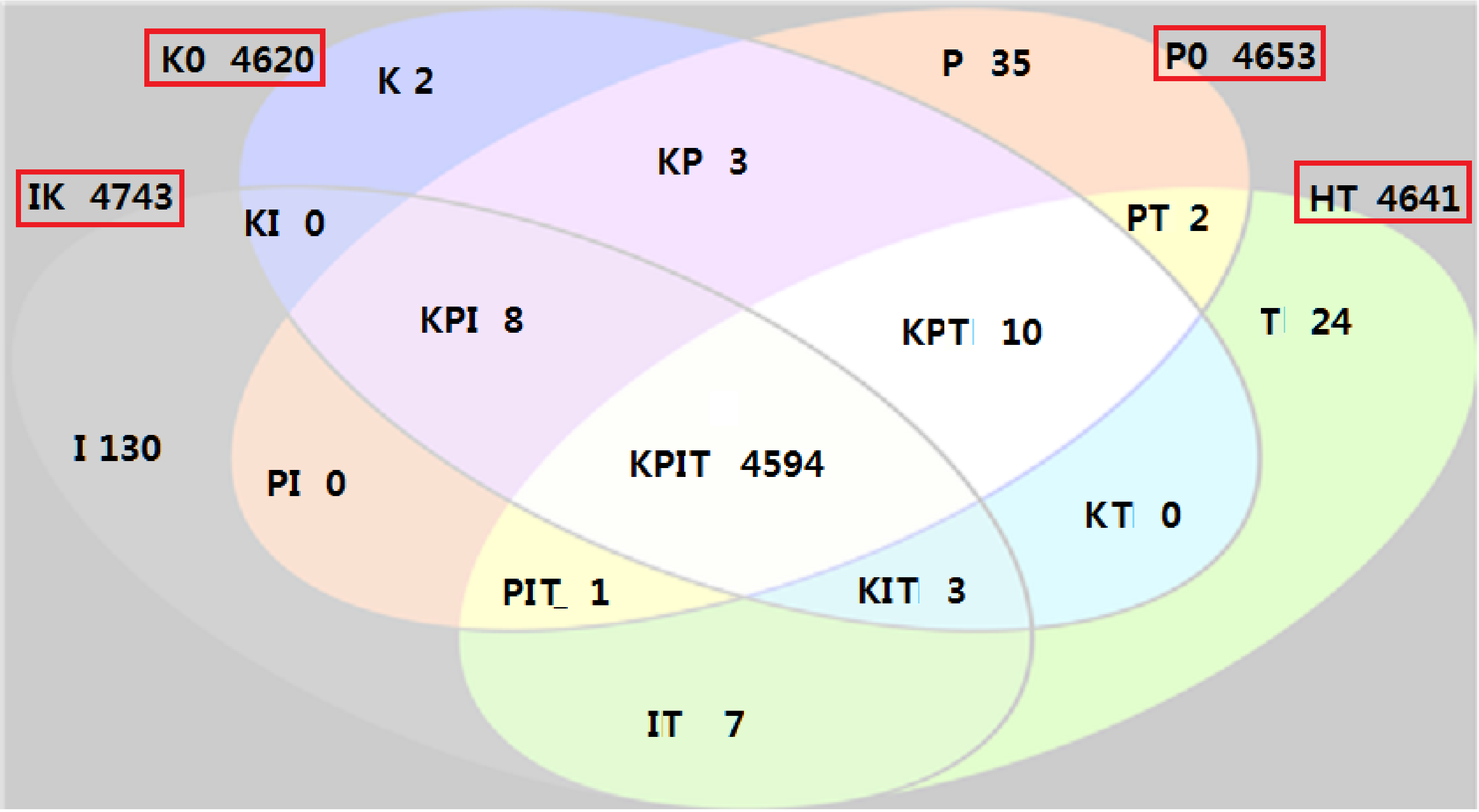
02.01 2016 — KGP formally formed

05 (?). 2016 — K-LGR v1.0

Next Steps

. Need to send translated meanings to CGP (in Mar/Apr (?), 2016)

Appendix. Hanja in K0, P0, IICORE/K, HT (Hanja Test)



Venn Diagram of 4 sets showing number of Hanja chars: (K-LGR v0.3, 2015.08.13.)

K0 (KS X 1001), P0 (KPS 9566), IK (IICORE: K), HT (Hanja Test) klgp168_2b_v03

Japanese GP (JGP) Update

9 March 2016

Hiro Hotta <hotta@jprs.co.jp>

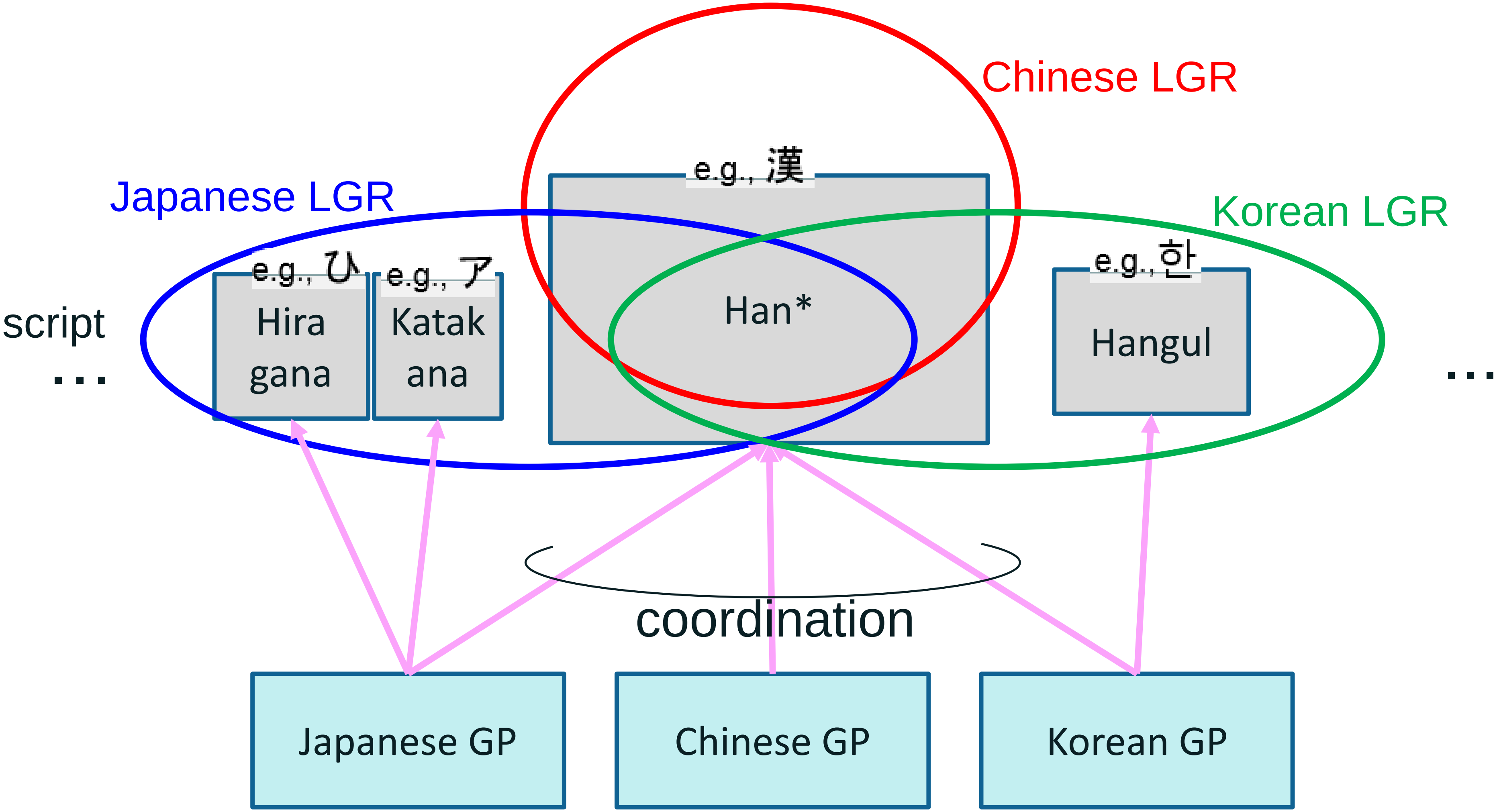
JGP Meetings & Related Events

- 2014
  - August 29 preparatory meeting (1)
  - September 12 preparatory meeting (2)
  - September 24 formal meeting (1)
  - October 24 formal meeting (2)
  - November 26 formal meeting (3)
  - December 18 formal meeting (4)
- 2015
  - January16 formal meeting (5)
  - February 4 formal meeting (6)
  - February 6 submission of JGP proposal to ICANN
  - February 20 formal meeting (7)
  - March 10 JGP establishment approved by ICANN
  - March 18 formal meeting (8)
  - April 15 formal meeting (9)
  - May 15-16 CJK coordination committee in Seoul
  - May 20 formal meeting (10)
  - June 17 formal meeting (11)
  - June 21-25 CJK coordination committee in during ICANN
  - September 29 formal meeting (12)
  - October18-22 CJK coordination committee in during ICANN

JGP Members

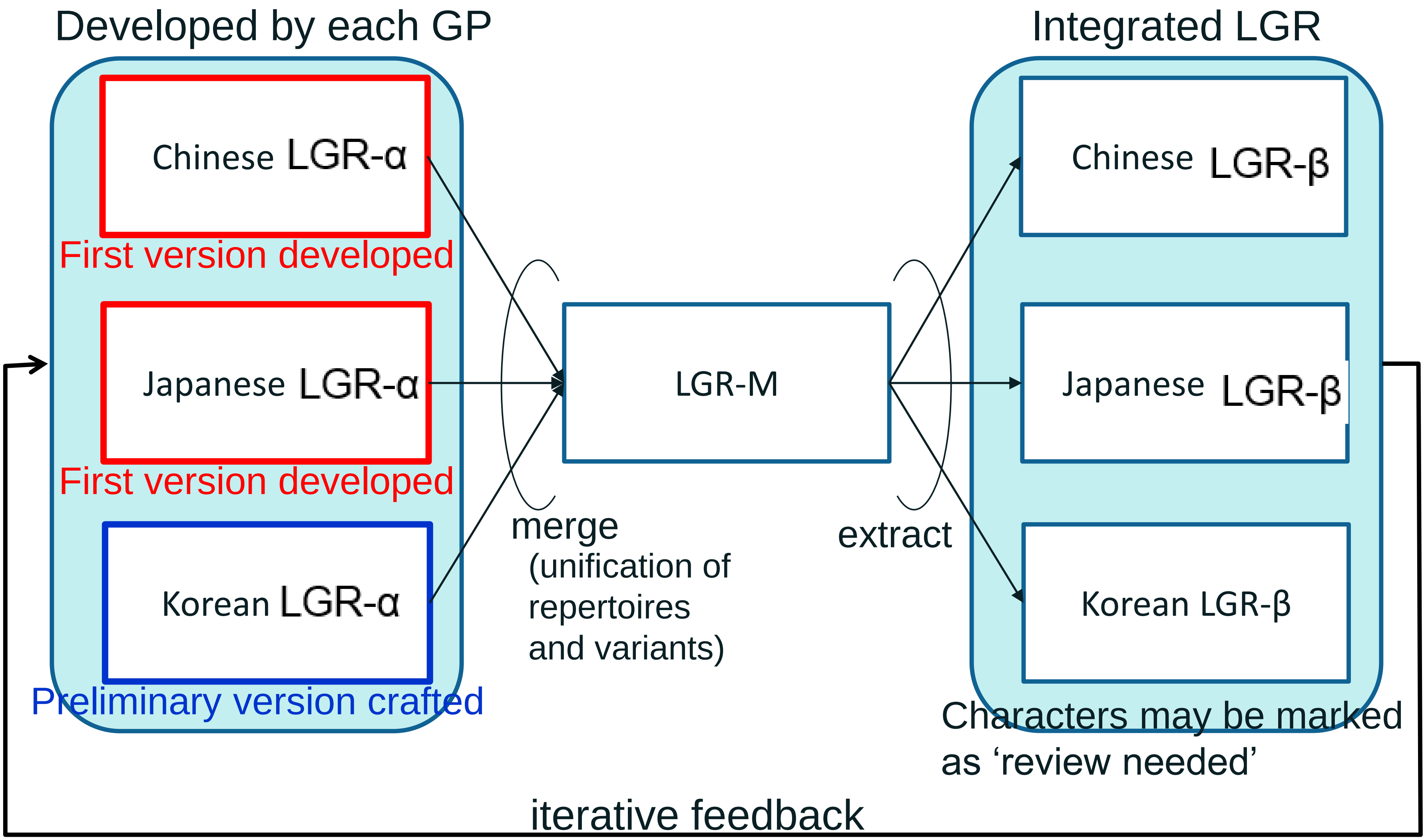
- Members and their expertise
  - Hiro Hotta Chair
    - Policy/business aspects of registry/registrar
  - Akinori Maemura Vice Chair
    - Internet governance and domain name in general
  - Shigeki Goto
    - Internet in general
  - Kazunori Konishi
    - Internet in general
  - Tsugizo Kubo
    - Trademarks and domain names
  - Yoshitaka Murakami (from February 4, 2015)
    - Trademarks and gTLD markets from registry/registrar perspective
  - Shuichi Tashiro
    - Character codes
  - Yoshiro Yoneya
    - Technical aspects of IDN, LGR

Relationship Among CJK Language LGRs



* “Han” is called “Kanji” in Japan, “Hanja” in Korea

Framework of CJK LGR Integration for Han Characters
(revised by agreement in Buenos Aires)



Developed by each GP

Chinese LGR-α

First version developed

Japanese LGR-α

First version developed

Korean LGR-α

Preliminary version crafted

merge
(unification of
repertoires
and variants)

LGR-M

extract

Integrated LGR

Chinese LGR-β

Japanese LGR-β

Korean LGR-β

Characters may be marked
as ‘review needed’

iterative feedback

Activities

- JGP establishment
  - Proposal submitted to ICANN (February 6, 2015)
  - Establishment approved by ICANN (March 10, 2015)
- Detailed task description of JGP
  - Done
  - Some more tasks or issues may come out or tasks may be modified on the way to LGR development
    - thorough discussion with ICANN/IP
    - thorough discussion with CGP and KGP (as well as IP)
    - thorough investigation inside JGP
- Development of Japanese LGR
  - CJK LGR integration procedure was developed, agreed, and revised by CGP, JGP, and KGP as a framework
  - As the input to the procedure, preliminary Japanese LGR (which is called LGR-α) was developed

Discussion Status for Japanese LGR- α

- Scopes of the character codes
  - Kanji, Hiragana, Katakana
  - For Kanji
    - JIS (Japanese Industrial Standard) level-1 and level-2
- Variants & each variant type
  - For Kanji
    - Japanese LGR-α will define no variants for itself
    - Integrated Japanese LGR (which is called LGR-β) will import (= passively adopt) variants of Chinese LGR-α and Korean LGR- α
    - Types of each variant in Japanese LGR will be defined in a systematic way == > need more investigation if reduction of the number of allocatable labels is needed, and how it can be done if needed
- WLE (whole label evaluation)
  - Japanese LGR-α may have no, or very limited number of, tiny rules even if defined == > need more investigation

Overview of Japanese LGR-α (J-LGR-α)

- Repertoire
  - Consists of characters from 3 scripts (Han, Hira and Kana – Jpan in ISO 15924)

Script	# of characters
Han	6358
Hira	85
Kana	89
Total	**6532**

- Variants & their types
  - No variants
  - Types of imported variants will be investigated and determined after LGR-α from CGP and KGP are proposed
- WLE
  - Rules (although not very many) are under discussion

Developments At & After Dublin (1)

- Is reduction of the number of allocatable labels really necessary?
  - Variant labels will exist by importing CGP variant characters, although JGP defines no variants
  - So, we analyzed Kanji domain names currently registered under .jp
    - Biggest size of the set of calculated variant labels will be 20,736
    - Biggest size of the set of variant allocatable labels as Japanese domain names will be 540
    - Biggest number of labels that are mutually variants registered under .jp is 4
  - IP (Integration Panel) has just requested JGP to reconsider the reduction of allocatable labels further

Developments At & After Dublin (2)

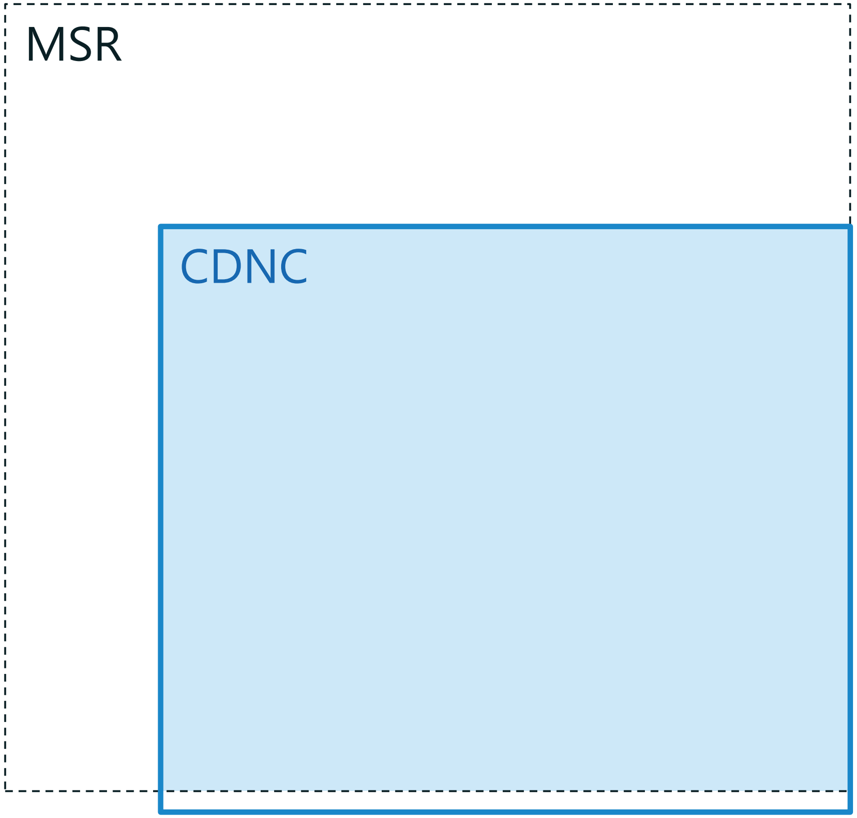
- Communication with the Japanese community
  - Presentation and discussion with Japan Trademark Association <Oct.2015>
    - No objection against draft LGR-α wαs raised
  - Presentation and discussion with various stakeholders at IGCJ (Internet Governance Conference Japan) event <Nov.2015>
    - No objection against draft LGR-α wαs raised

- Related activity
  - Submission of public comments to “Guidelines for Developing Reference Label Generation Rule sets (LGRs) for the Second Level” <Jan.2016>

Chinese GP Update

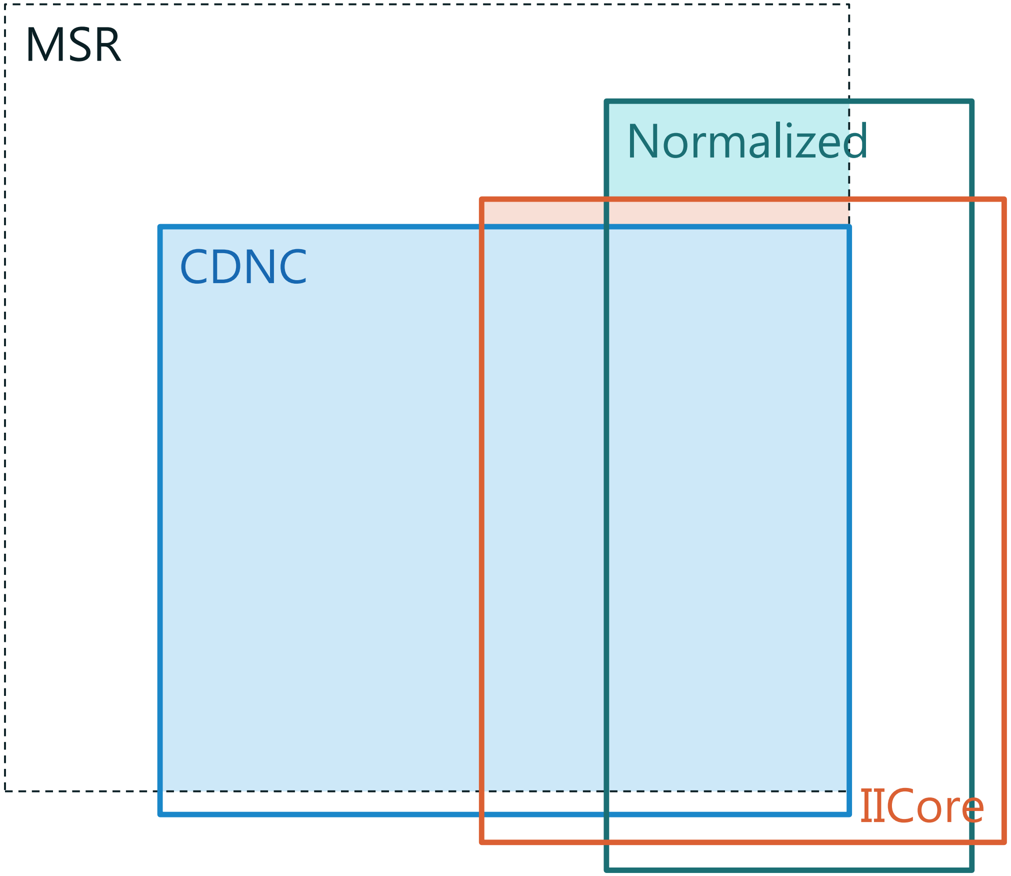
Wei WANG & Kenny HUANG | IDN LGR Workshop | 9 March 2016

CGP Repertoire and CDNC2015



- In July 2015, CDNC Taiwan meeting urged to add all CDNC chars into CGP repertoire, to reach consistency between the CDNC SLD operation and future TLD operation.
  - Compared with original CDNC/CN/TW table (19520 chars) http://www.cdnc.org/gb/research/file/CDNC_unicode.txt
  - **CDNC Table 2015** has 41 new chars requested by HK community http://www.cdnc.org/gb/research/file/unicode.txt 39 of which fall in the range of MSR.

CGP Repertoire=CDNC2015+Normalized+IICore



- **The Normalized**

  China's State Council published Normalized Hanzi List for Common Use in 2013,

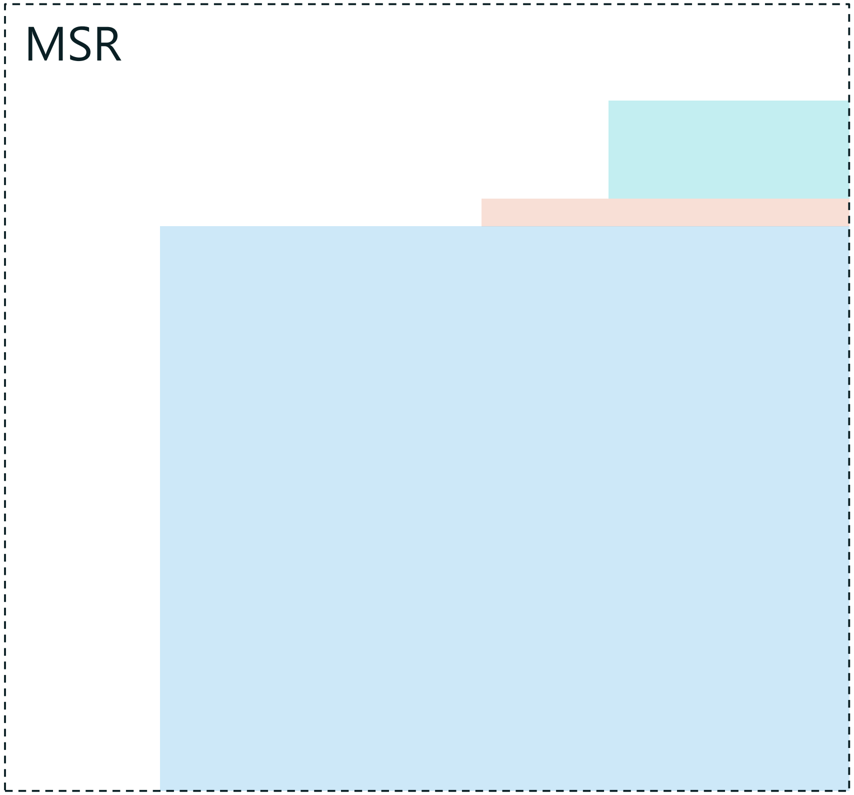
  27 new chars in the range of MSR

- **IICore**

  International Ideographs Core

  145 new in the range of MSR

CGP Variants



- CDNC and CGP linguists analyzed 172 new chars from the Normalized and IICore. In particular, linguists studied 107 chars which are also in JGP repertoire and updated variant setting of CGP repertoire.

Example:

麹 (9EB9)	Non CDNC2015	Normalized	IICore	JGP

Simp	Trad	Var

曲(66F2)	曲(66F2)
	麴(9EB4)
		麹(9EB9)曲(66F2)麴(9EB4)
		麯(9EAF)

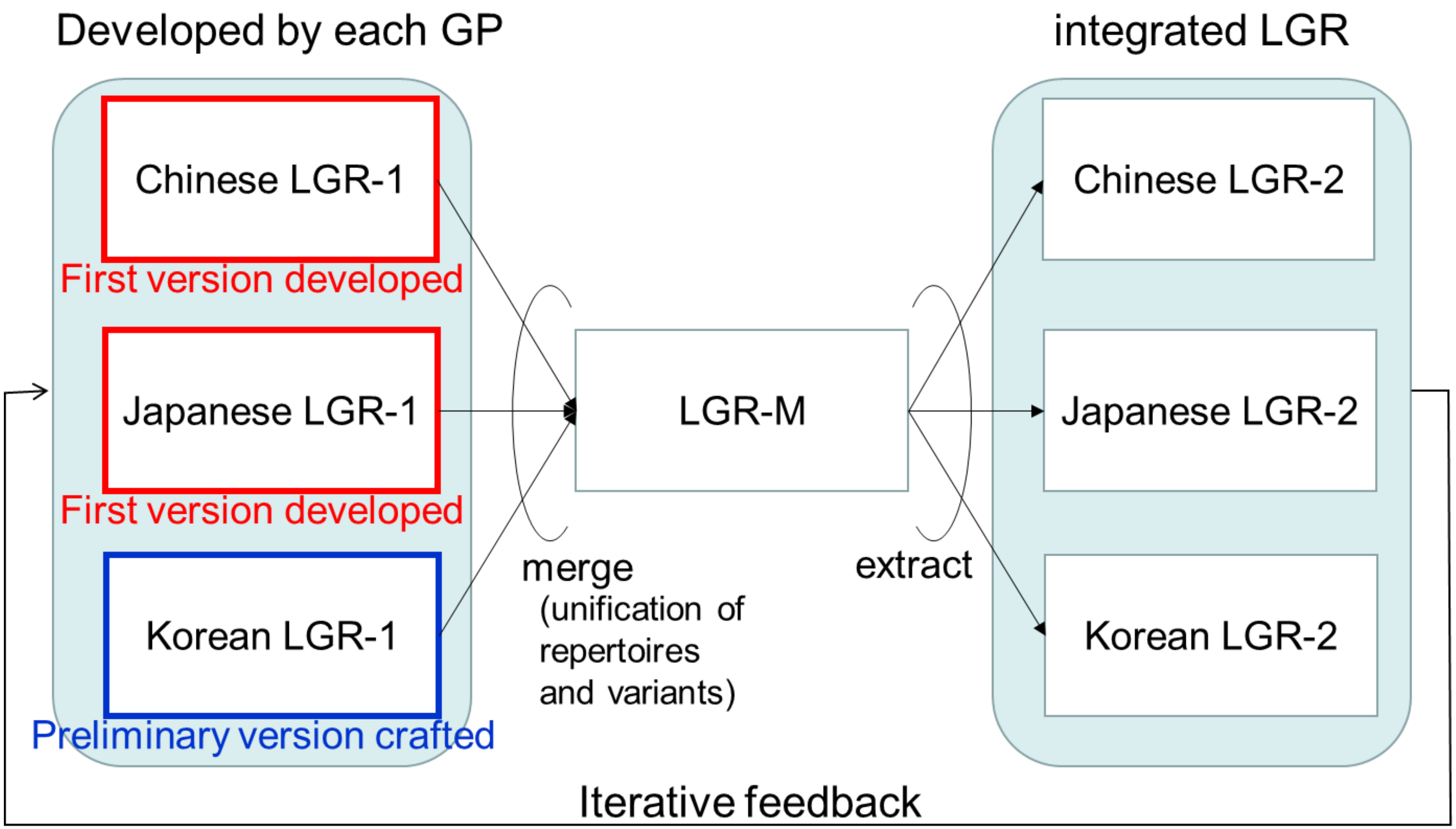
CGP LGR XML

- CGP fixed the flaws in the last version of CGP LGR XML document:
  - Complete the description of Reference ID
  - Add the type of “reflective” (r-simp, r-trad, r-both)
  - Correct spelling error of “block”
  - Add comments to the action rules

CGP Repertoire vs JGP and KGP

- Besides 107 JGP chars mentioned above, CGP does not seek to borrow more chars or variants from JGP.
- KGP provided K-LGR v0.3 in September, including 4819 Hanja char, all falling in the range of CGP repertoire.

The Next Step



CJK agreed to generate merged LGR based on the algorithm proposed by Yoneya San.

Challenge

- It still needs further in-depth exchange and compromise to reach the consensus on CJK variant set
  - “机”and“機” have different meanings in Japanese language environment
  - Similarly, K listed 258 unacceptable variant groups in C LGR

C has invited J and K to visit Beijing at 20~21, March.

Thanks

Q&A