
ICANN
COMMUNITY FORUM
64
KOBE
9–14 March 2019

IDN Root Zone LGR Workshop

IDN Program, ICANN

ICANN64
13 March 2019

Overview of Session Presentations

- Variant Considerations - Michel Suignard
- Update by RZ-LGR Study Group - Dennis Tanaka
- Community Updates
  - Chinese GP Update - Wei Wang, Kenny Huang
  - Japanese GP Update - Yoshiro Yoneya
  - Korean GP Update - Dongman Lee
- Q/A

Variant Considerations

Michel Suignard
Member, Integration Panel

Variant Relation

- Variants are EXCHANGEABLE: Users will accept one for the other
- Variants are symmetric:
  - A ~ B means B ~ A
- Variants are transitive:
  - A ~ B and B ~ C means A ~ C

Variant Types

- Variants can be semantically equivalent, phonemically exchangeable, functionally exchangeable, visually identical
- Semantic variants:
  - Chinese: Traditional versus Simplified ideographs 萬 万
  - Arabic: Orthography variants ك ک
- Phonemic variants:
  - Ethiopic (Amharic writing system only) ሆ ሐ ኆ
- Functionally exchangeable variants:
  - Latin (ligature œ versus the sequence ‘oe’)
- Visually identical variants
  - Latin (schwa ‘ə’ versus turned e ‘ɘ’)
  - Arabic (shared positional forms, stylistic variants)

Code Point Variants Versus Variant Labels

- LGRs define variants among code points (or sequences)
  - These result in variant labels after dispositions are applied
- Users interact with variant labels
  - Code point~~s~~ variants that cannot be used to create a variant label do not need to be defined.
    - Example: Combining marks unless the base characters have also variants
  - Code points have no context other than the label.
    - Example: Most Devanagari users don’t expect a Nukta behind a Vowel (therefore creation of a variant pair with the naked vowels)

In-Repertoire and Out-of-Repertoire Variants

- Variants are not restricted to being in-repertoire (or even in-script), they can be out-of-repertoire and cross-script
- GPs must investigate all in-repertoire and declare all of them in their LGRs
- Out-of-Repertoire or cross-scripts can be added through integration of other LGRs
- GPs cannot add cross-script variants that would induce in-script variants in another script without agreement of the other GP (as a consequence, it is not possible to force an in-script variant in a script already included into the latest RZ-LGR)
  - Exception for shared script scenario
- No cross-script variants can create variants between ASCII letters or sequences

Special Cases of Generic Shapes

- Generic, simple shapes lack identifying features
  - High risk example: .ooo is a delegated non-IDN gTLD

Script	.ooo	.coco	.olol
Latin	.ooo	.coco	. olol
Cyrillic	.ooo	.coco	. olol
Greek	.ooo		
Armenian	.ooo		
Oriya	.OOO		.OlOl
Malayalam	.OOO		
Myanmar	.ooo	.coco	

Determining Visually Exchangeable Variants

- It is not mere visual similarity, such cases are resolved outside of LGRs
- But "*Even for variants based on visual similarity, there exists a subset of evaluation rules that could be applied in an automated manner, obviating the need for further case-by case or even contextual review*" [Procedure A3.3]
- IDN labels are much more restrictive than normal text
- Identifiers are not restricted to being actual words
- Procedure and IAB have directed the process to be biased to the conservative side
  - A TLD cannot be visual identical to another one because users will be misled

In-Script Exchangeability

Example of TABERU	
食べる	Japanese "to eat", Kanji + Hiragana
食べる	Substituting Katakana ベ

Exchangeable or not?

Allocatable and Blocked Variants

- "*The output of this procedure should aim to maximize the number of blocked variants, and to minimize the number of allocatable variants*." [procedure A3.3]

- Therefore, all LGRs must contain mechanism to reduce the number of allocatable variants when such a mapping/action set is defined.

- Examples of mitigations:
  - Arabic, using WLE rules to limit permutations
  - Chinese (Draft), using a mix of mapping types and actions to reduce the number of allocatable labels (typically: original, 1 traditional, 1 simplified)
  - Tamil SRI/SHRI ஶ்ரீ / ஸ்ரீ alternation, using WLE rules to restrict permutation in a single label

Variants Are Defined as Mapping + Context

Variant 1	Glyph	Variant 2	Glyph		Type	Required Context	Comment
0906	आ	**0906 093C**	आ़	↔	**blocked**	**not: followed-by-Nukta**	**Devanagari variant**

- Example is a NULL Variant because 093C maps to *nothing* when mapping from 0906 093C (Variant 2) to 0906 (Variant 1)
- Variant 1 to Variant 2 relationship does not exist if the Variant 1 character is followed by a Nukta (U+093C) to avoid possible recursion:
  - The 0906 in the new variant mapping including it could be mapped again to keep making the variant mapping longer
  - This context rule is required to keep the variant set transitive.
- Variant 2 to Variant 1 relationship seemingly does not require the same context because the string 0906 093C 093C is invalid per Nukta context rule.
  - However it is required to keep the variant mappings symmetric

References

- 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

- Guidance on Designing Label Generation Rulesets (LGRs) Supporting Variant Labels
  https://tools.ietf.org/html/rfc8228

- Root Zone Label Generation Rules 2.0
  https://www.icann.org/sites/default/files/lgr/lgr-2-overview-26jul17-en.pdf

- Out of Repertoire Variants in Root-Zone LGR and Proposals
  https://community.icann.org/download/attachments/43989034/Out-of-Repertoire-Variants_2017-09-25a.pdf

Update by RZ-LGR Study Group

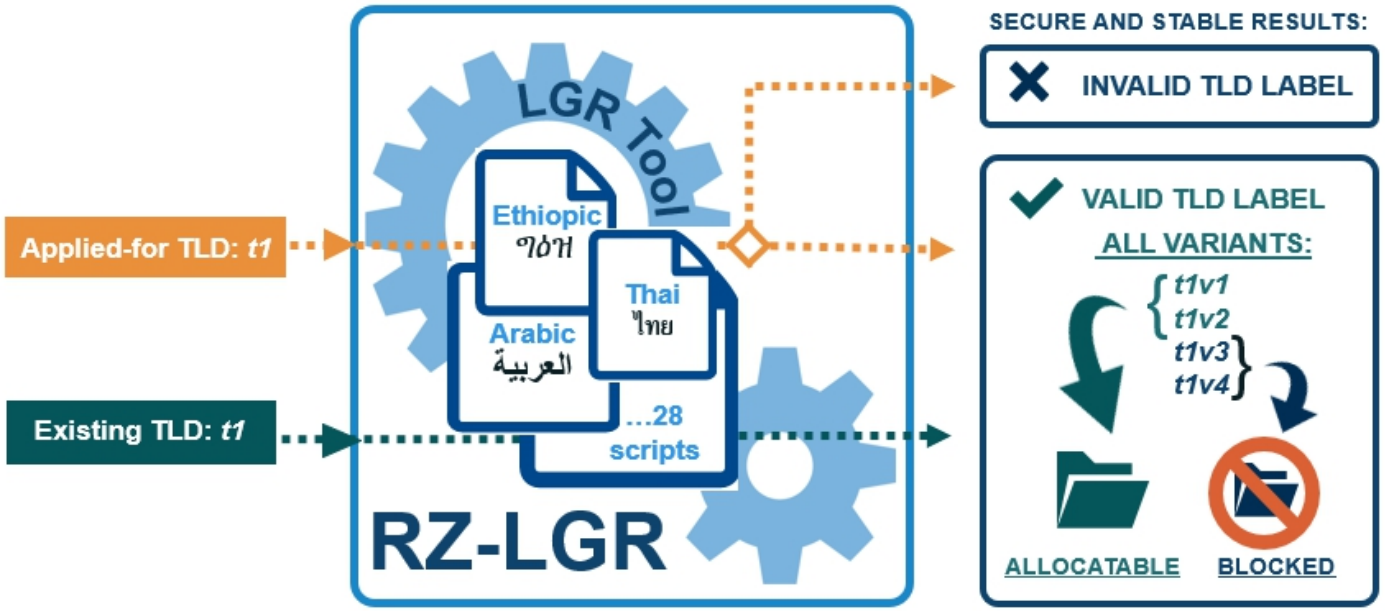
Dennis Tanaka
Chair, RZ-LGR Study Group

Agenda

- Background
- Scope of Work
- Current status
- Next Steps

Background

- RZ-LGR available through the LGR Procedure
  - Several scripts already integrated; many others in-progress
- Need of a harmonized way to use the RZ-LGR for ccTLDs and gTLDs
  - Single source to validate top-level labels and calculate variant labels
- Need for a technical assessment of the implementation of the RZ-LGR
  - Technical considerations for subsequent policy



Background – Study Group Members

	Name	Organization	Sponsoring Organization
1	Mirjana Tasic	.rs and .срб	ccNSO
2	Edmon Chung	.asia	GNSO
3	Gaurav Vedi	Dominion Registries	GNSO
4	Dusan Stojicevic	Gransy	GNSO
5	Dennis Tan Tanaka	Verisign	GNSO
6	Wei Wang	KNET	GNSO
7	Ajay Data	XGENPLUS	GNSO
8	Alireza Saleh	IRNIC	IAB
9	Dessalegn Yehuala	Addis Ababa Univ. and Ethiopic Generation Panel	
10	Harsha Wijayawardhana	Sinhala Generation Panel	

Scope of Work

1

WHO will use it?

- TLD applicant (ccTLD, gTLD)
- Generation and Integration Panels
- Other stakeholders

2

WHAT does it do?

- Syntax validation
- Calculation of variant labels and disposition values
- What if RZ-LGR calculation is not accepted?

3

WHY is it important?

- Single source and/or repository, for consistency and predictable results
- But, what about scripts not yet integrated in the LGR? What are the technical issues subsequent policy would need to address

4

WHEN do you apply it?

- Existing TLDs and new TLD applications
- gTLDs: application window
- ccTLDs: Fast Track process (rolling)
- Reserved TLD labels

5

WHERE do you find it?

- Implementation (i.e, specs, test cases)
- Maintenance (e.g., update to repertoire, variant rules, etc.)
- Repository of normative XML

6

Other Considerations

- Variant states and transition among states
- Limits on allocatable variant labels
- Other security and stability considerations (e.g., single character IDN TLDs)

Not in Scope

- Semantic validation
  - IDN ccTLD, Geo-Names, Brand, Community, etc.
- Limiting number of allocatable variant TLDs
- How to process TLD applications whose script is not yet supported by the Root Zone LGR.

Proposed Recommendations (1/2)

- RZ-LGR is meant for all forms of top-level domain names (u-labels, a-labels and all other ldh labels)
  - ASCII set (a-z) is a subset of Latin script; cross-script variants in Cyrillic, Greek or Armenian scripts are possible.
- For scripts or writing systems integrated into the RZ-LGR, the RZ-LGR is the sole authoritative source
  - to validate top-level domain labels, and
  - to calculate variant labels and disposition values.
- Policy should not overturn calculations of the RZ-LGR; doing so would invalidate the entire RZ-LGR.
- For scripts not yet incorporated into the RZ-LGR, the SG defers to a policy development process to determine whether it is advisable to delegate an applied-for label which can't be validated by the RZ-LGR.

Proposed Recommendations (2/2)

- Changes to the repertoire, variant code points or whole label evaluation rules should not affect existing top-level domains.
  - Subsequent changes to components of the RZ-LGR must conform to the LGR Procedure.
- There should be one and only one authoritative source for the RZ-LGR xml file (e.g. IANA).
- ICANN should make available a non-authoritative implementation of the RZ-LGR as a community service.
- Number of allocatable variant labels should be as small as possible.
- ICANN and the relevant SOs needs to develop a process to resolve cases when a TLD applicant does not agree with the calculations of the RZ-LGR
  - Any resolution that requires changes to the RZ-LGR must conform to the LGR Procedure

Thank You

Update by Chinese Generation Panel

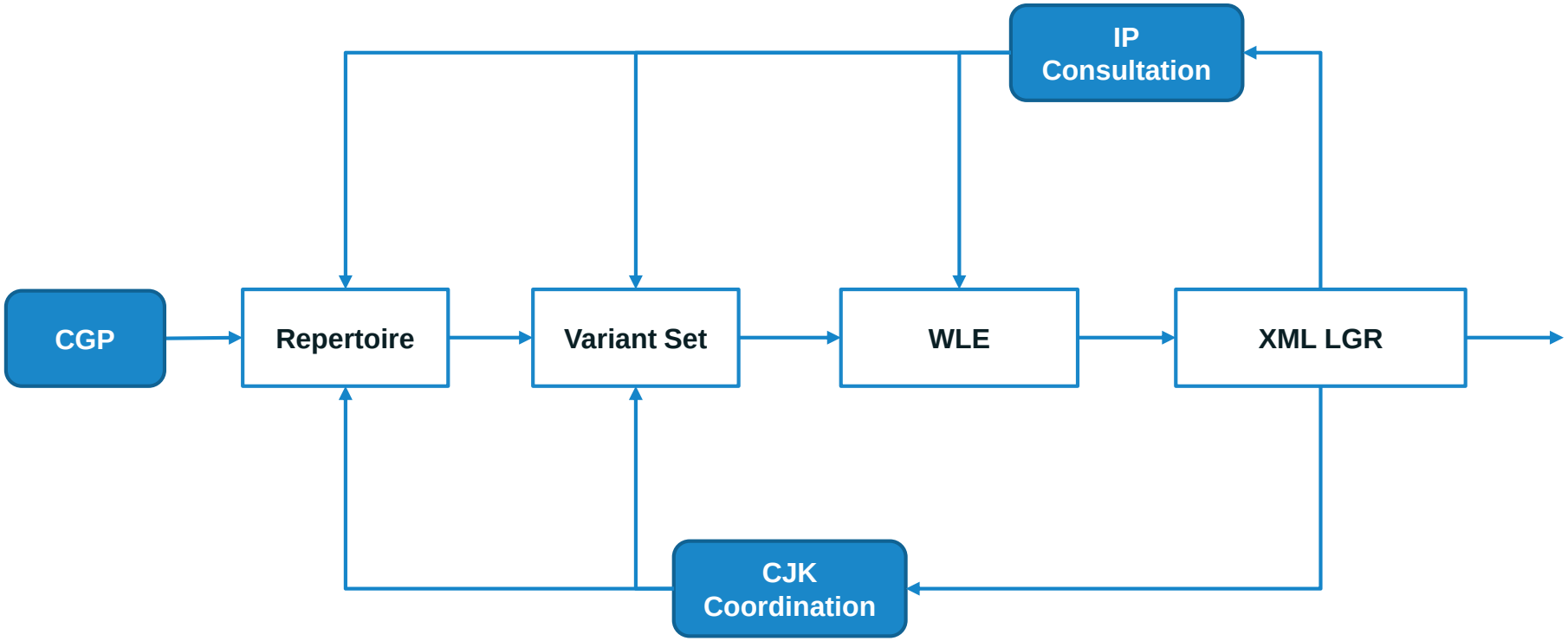
Wei Wang
Kenny Huang
Co-Chairs, Chinese GP

Agenda

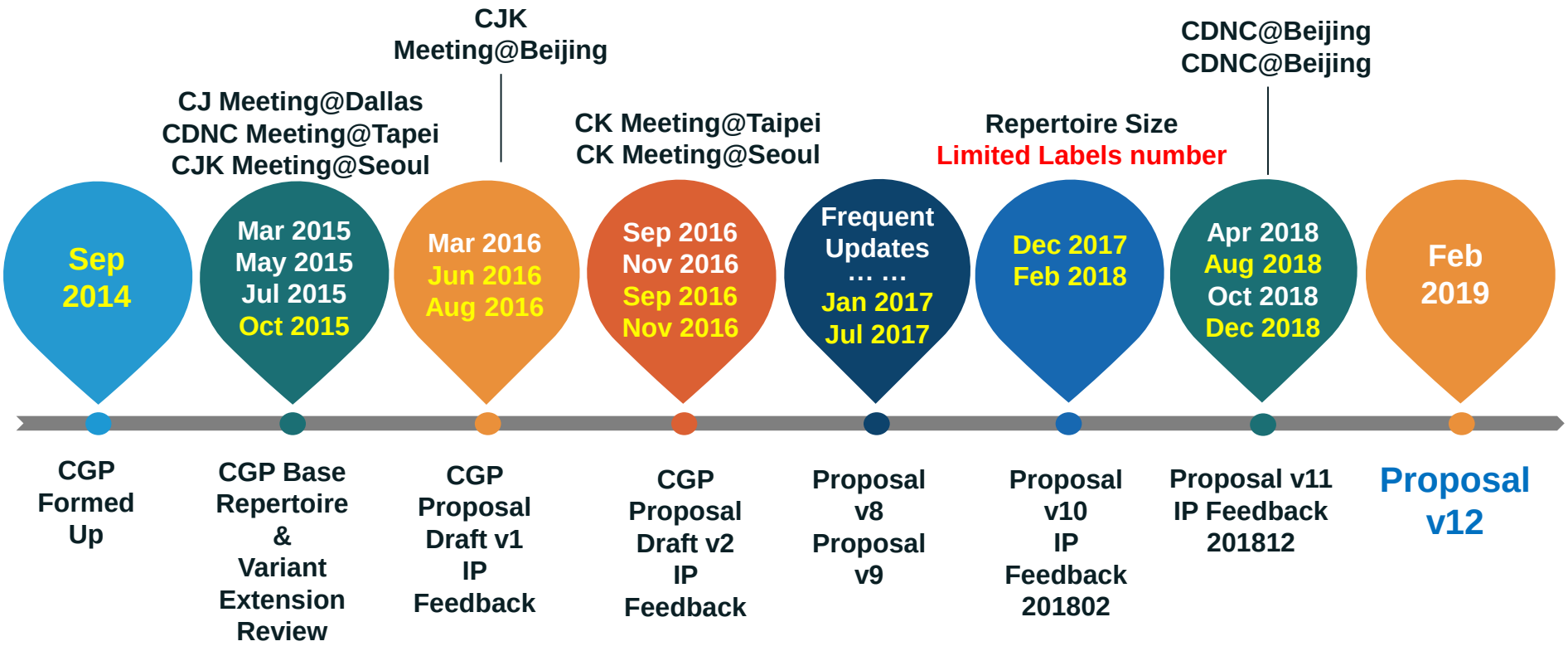
1. CGP Work Overview
2. CJK Coordination
3. CGP Proposal Draft 201902
4. Visual Similarity
5. Next Step

1 CGP Team and Work Process

- Members, 23 experts from 10 countries/regions
  - China mainland, Taiwan, Hong Kong, Macau, Singapore, Malaysia, as well as members from Europe and North America.
- Advisor, Edmon Chung
  - CEO of dotAsia and Co-Chair of the Universal Acceptance Steering Group
- CJK coordination working group



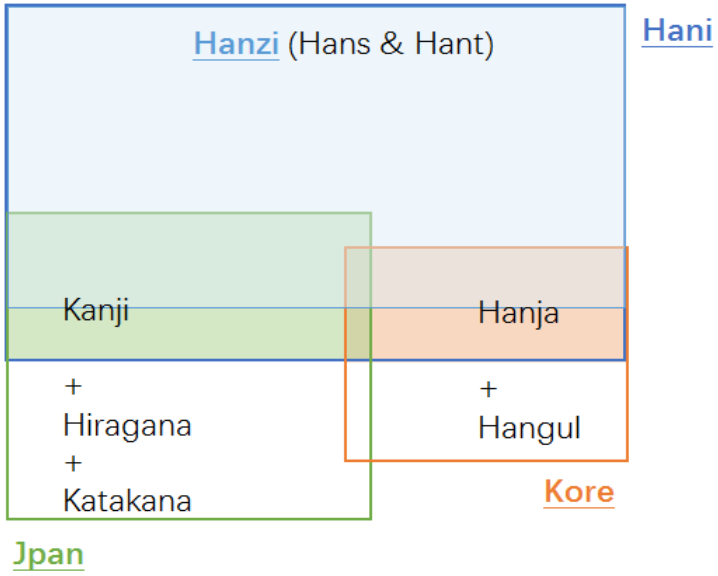
1 CGP Work Review



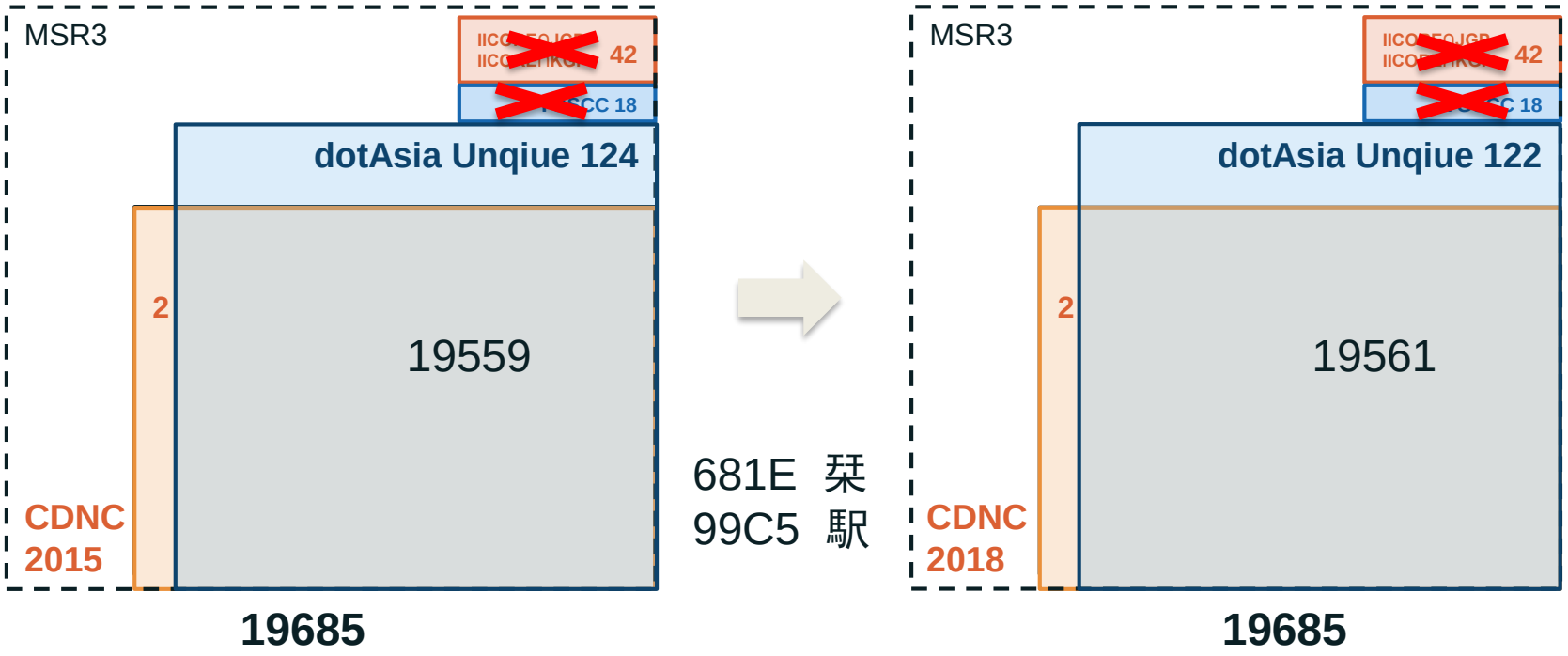
2 CJK Coordination

- Script and Languages Covered

Language	ISO 15924 Code	Countries	Local Names of the Script
Chinese	cdo, cjy, cmn, cpx, czh, czo, gan, hak, hsn, lzh, mnp, nan, wuu, yue, zho	China	汉字 Hanzi
Japanese	jpn	Japan	漢字 Kanji
Korean	kor	Korea	한자 Hanja

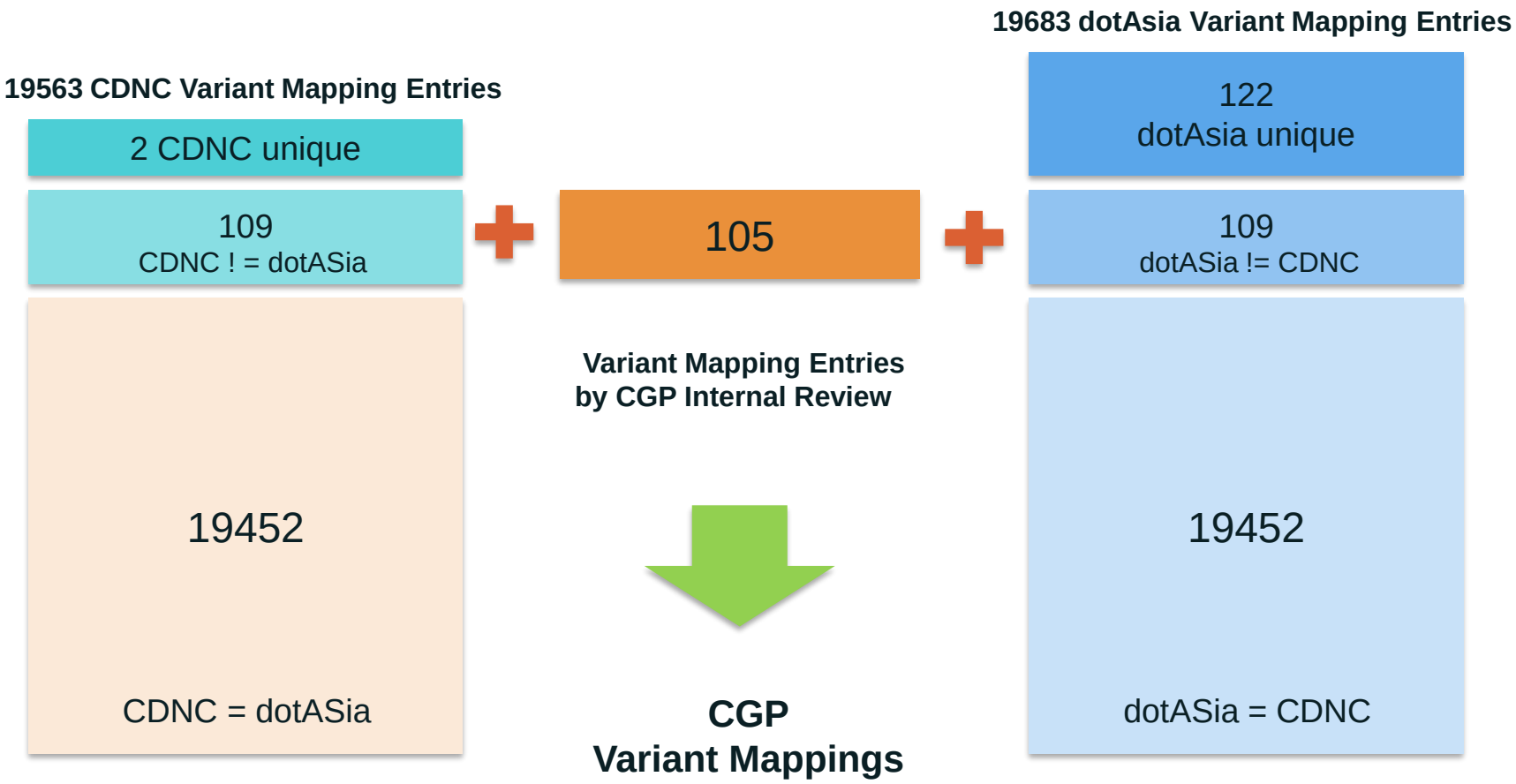


CGP character repertoire and source



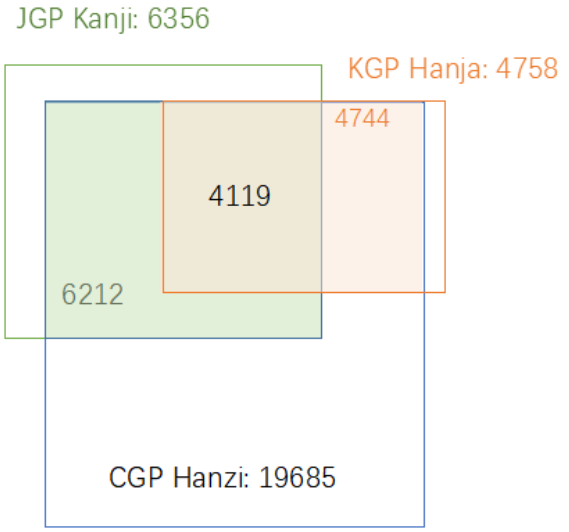
Coordination within CGP

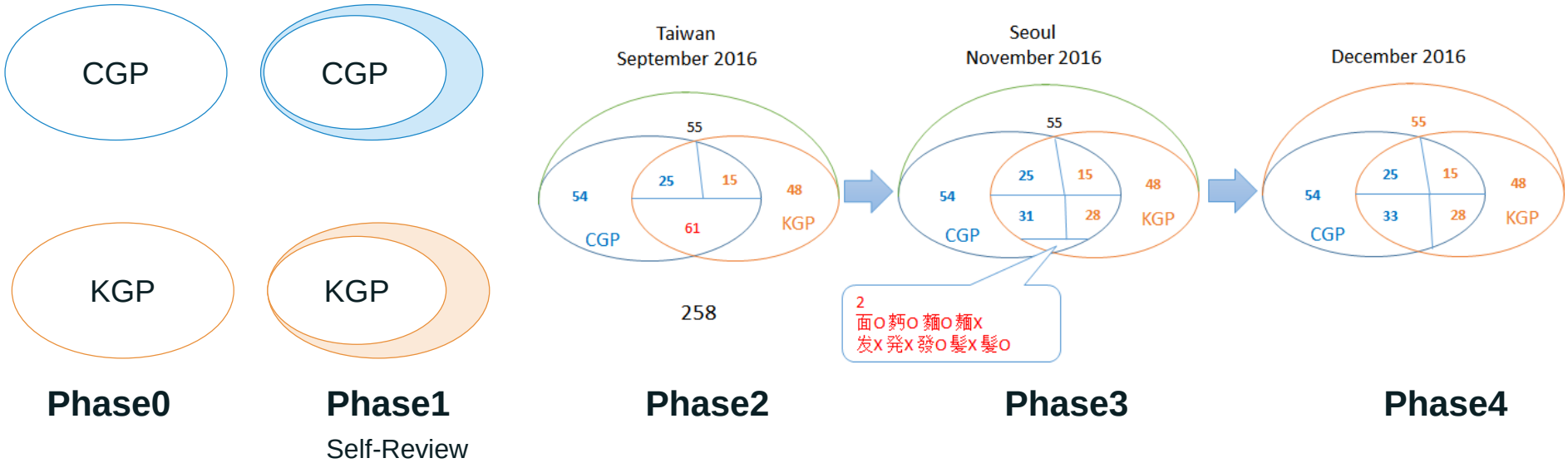
- Repertoire = CDNC + dotAsia
- Variant Mappings = CDNC + dotAsia + CGP Internal Review



2 CJK Coordination

- Coordination between C, (J) and K
  - 445 variant mapping entries
  - 146 unacceptable variant groups





Reconsider C-K Coordination from a conservative perspective

Statistics of disputed groups in .CN/TW/HK/网址 registration database

C Keep	Type1	Number of labels containing each disputed variant char > 0; Semantic meaning of variant char in the labels are the same; CGP would insist to KEEP them as variants	97	141
	Type2	Number of labels containing one disputed variant char ≈ 0; Semantic meaning of variant char in the labels are the same; CGP would insist to KEEP them as variants	44	
C Drop	Type3	Number of labels containing each disputed variant char ≈ 0; CGP would DROP the variant mappings and split them into independent characters	12	117
	Type4	Number of labels containing one disputed variant char ≈ 0; CGP would DROP the variant mappings and split them into independent characters	67	
	Type5	Number of labels containing any disputed variant char != 0; But semantic meaning of the chars in the labels NOT the same; CGP would DROP the variant mappings and split them into independent characters	38	

CK Coordination
... ...
146 Groups
445 entries

∩ 117 Groups

CK Coordination
... ...
117 groups
327 entries

4 Visual Similarity

Unicode consortium‘s confusables list
https://www.unicode.org/Public/security/11.0.0/confusables.txt

Source	Glyph	Target	Glyph
53E3	口	56D7	口
571F	土	58EB	士
58AB	壿	58FF	壿
676E	柿	67FF	柿
8D7F	越	8D86	越
9E42	鶒	9E43	鶒

Disposition Pinciple:

Some will be kept unrelated with explaination

-- 571F土 & 58EB士、9E42鶒 & 9E43鵲

Non-moden used ones may be treated as visual identical variants

-- 58AB壿 & 58FF壿、676E柿 & 67FF柿、8D7F越 & 8D86越

Radical may be treated as visual identical variants

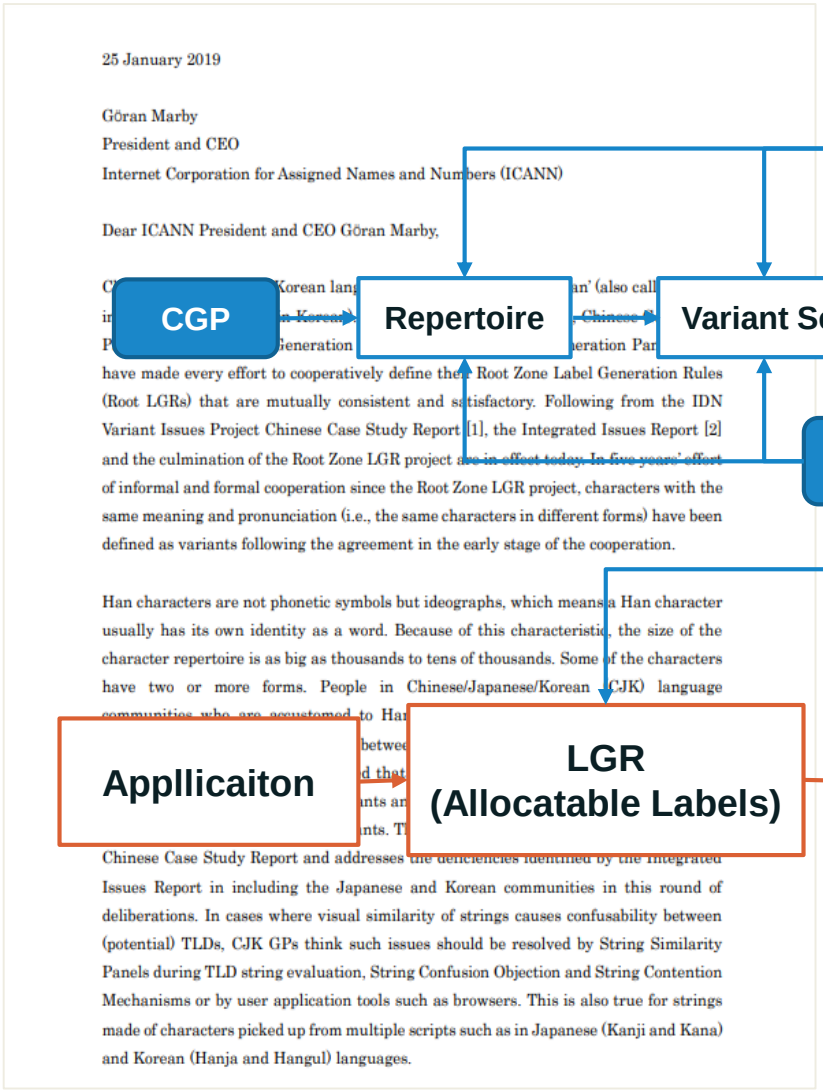
-- 56D7 口

Source	Glyph	Target	Glyph
C2A5	슩	4ECA	今
C2B4	슴	5408	合
C4F0	쓰	4E1B	丛
B9C8	마	535F	卟
B258	뉘	723F	爿

Source	Glyph	Target	Glyph
U+3078	ヘ	U+30D8	ヘ
U+30AA	オ	U+624D	才
U+30AB	カ	U+529B	力
U+30ED	ロ	U+53E3	口
U+30CF	ハ	U+516B	八
U+30C8	ト	U+535C	卜
U+30CB	ニ	U+4E8C	二
U+30A8	エ	U+5DE5	工

4 Visual Similarity

https://www.icann.org/en/system/files/correspondence/wei-et-al-to-marby-25jan19-en.pdf



25 January 2019

Göran Marby
President and CEO
Internet Corporation for Assigned Names and Numbers (ICANN)

Dear ICANN President and CEO Göran Marby,

have made every effort to cooperatively define their Root Zone Label Generation Rules (Root LGRs) that are mutually consistent and satisfactory. Following from the IDN Variant Issues Project Chinese Case Study Report [1], the Integrated Issues Report [2] and the culmination of the Root Zone LGR project are in effect today. In five years' effort of informal and formal cooperation since the Root Zone LGR project, characters with the same meaning and pronunciation (i.e., the same characters in different forms) have been defined as variants following the agreement in the early stage of the cooperation.

Han characters are not phonetic symbols but ideographs, which means a Han character usually has its own identity as a word. Because of this characteristic, the size of the character repertoire is as big as thousands to tens of thousands. Some of the characters have two or more forms. People in Chinese/Japanese/Korean (CJK) language

Chinese Case Study Report and addresses the deficiencies identified by the Integrated Issues Report in including the Japanese and Korean communities in this round of deliberations. In cases where visual similarity of strings causes confusability between (potential) TLDs, CJK GPs think such issues should be resolved by String Similarity Panels during TLD string evaluation, String Confusion Objection and String Contention Mechanisms or by user application tools such as browsers. This is also true for strings made of characters picked up from multiple scripts such as in Japanese (Kanji and Kana) and Korean (Hanja and Hangul) languages.

5 Next Step

- Sync up CNDC / dotAsia IDN table with CGP
- More conservative and secure solution for C-K Coordination
- Visual Similarty
  - Option1: Solve the issue in string evaluation process
  - Option2: Solve all 6 Hanzi-Hanzi pairs within CGP LGR
  - Option3: Wait J and K to solve Kanji-Hanzi and Hangual-Hanja pairs

Update by Japanese Generation Panel

Yoshiro Yoneya
Member, Japanese GP

Agenda

1. Steps of Japanese Generation Panel (JGP)
2. Generation Panel Membership
3. Overview of the Japanese Root LGR
4. Coordination among Chinese/Japanese/Korean GPs
5. Reduction of allocatable variant labels
6. Handling characters with visual similarity

1 Steps of Japanese Generation Panel (JGP)

- Mandate
  - Proposing LGR for TLDs of Japanese language/scripts that co-exists in harmony with LGRs for other languages/scripts
- Steps

Step1 : Populate JGP with diverse experts (Aug/2014~)

Step2 : Define the requirements and basic framework of Japanese LGR based on the expertise and experience of Japanese IDNs (~April/2015)

Step3 : Coordinate with other language Generation Panels whose languages interrelated with Japanese (~February/2017)

Step4 : Finalize LGR following necessary consultation with IP and Japanese community (March/2017~)

We are still here!
Rather, we are requested by ICANN/IP to go back to Step2

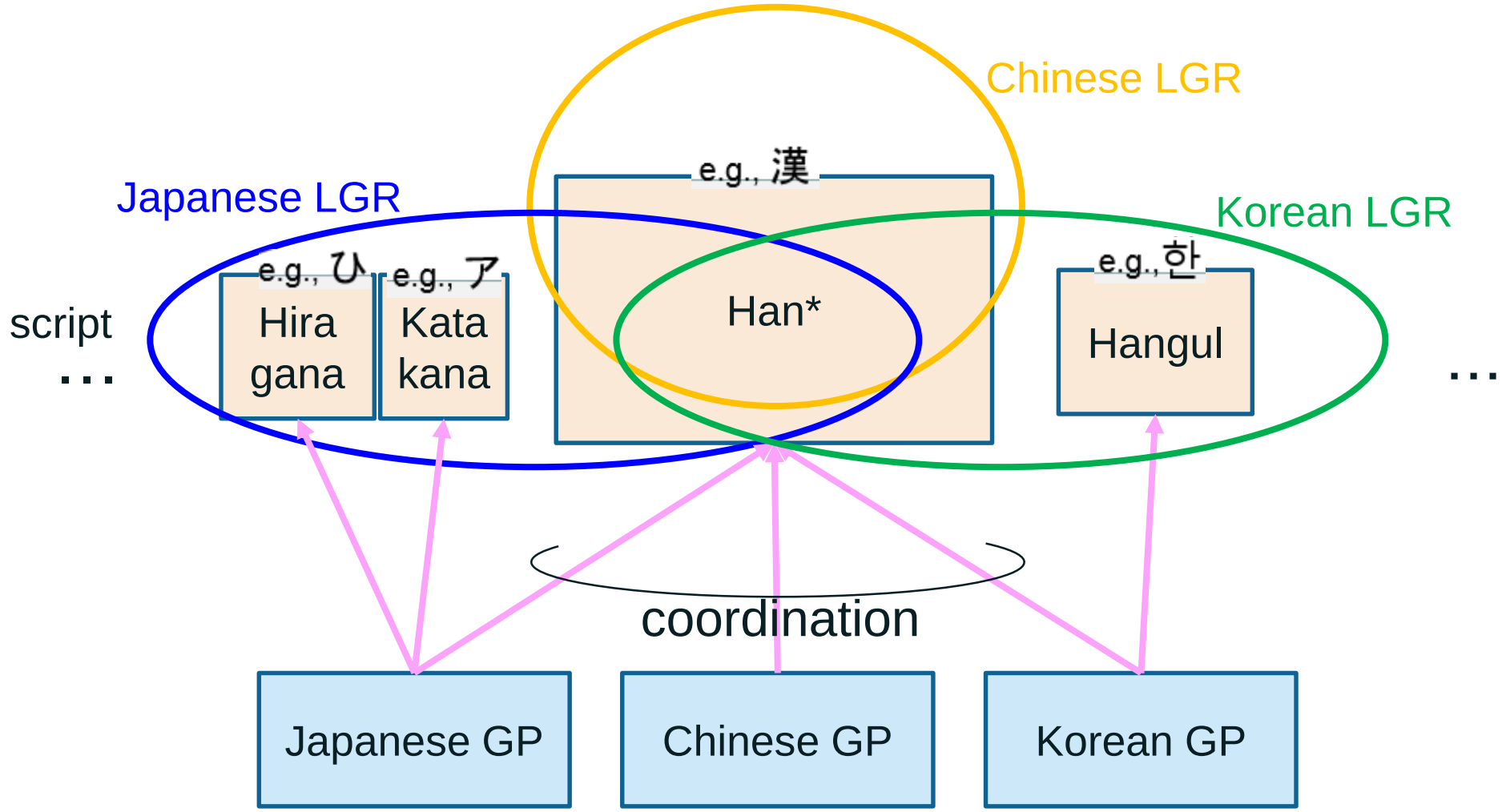
2 Generation Panel Membership

- **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**
  - Trademarks and gTLD markets from registry/registrar perspective
- **Shuichi Tashiro**
  - Character codes
- **Yoshiro Yoneya**
  - Technical aspects of IDN, LGR

3 Overview of the Japanese Root LGR

- Scopes of the character codes
  - Kanji, Hiragana, Katakana (can be mingled in a label)
    - JIS (Japanese Industrial Standard) level-1 and level-2
    - >6,000 characters
- Variants
  - For Kanji
    - Japanese LGR will define no variants for itself
    - Final Japanese LGR will import (= passively adopt) variants defined by Chinese LGR and Korean LGR
  - For Hiragana, Katakana
    - No variants
  - No variants for visual similarity : still unresolved
- WLE (whole label evaluation)
  - Rules for reduction of allocatable variants

4 Coordination: Chinese/Japanese/Korean GPs



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

Coordinated definition of variants has been completed (~February/2017)

5 Reduction of Allocatable Variant Labels

- Any combination of characters is allowed in Japanese labels as in the case of Japanese words in daily life
- The above may make the number of variant strings very huge, considering that many variant groups are imported from Chinese and Korean Root LGRs
  - E.g., 慶応大学 has 3 variant strings – 慶應大学/慶応大學/慶應大學
- Reduction of the number of allocatable variant labels was requested by ICANN/IP to prevent the explosion of root zone size
- With IP's suggestion, JGP solved it by limiting allocatable strings by employing the notion that "allocatable labels consist of daily-use (Joyo) Kanji
  - It reduces the maximum number of allocatable labels of an actually registered Japanese label under .JP from 486 to 8

6 Handling Characters with Visual Similarity

- Initial design of Chinese/Japanese/Korean Root LGRs is to define variants as characters with the same meaning and pronunciation having different forms (i.e., different versions of the same character)
  - Despite the above, initial design of Japanese Root LGR has no variants, except those variant definitions imported from Chinese and Korean LGRs
- However, ICANN/IP started to request Chinese/Japanese/Korean GPs to handle characters with visual similarity in their Root LGRs (March/2017)
- Chinese/Japanese/Korean GPs decided to formally request ICANN to withdraw such request
  - Correspondence to Göran Marby on 25 January 2019
  - Response from Cyrus K. Namazi on 15 February 2019
    - Not yet resolved

The Correspondence: Abstract

- CJK GPs believe that incorporating variants into LGR in order to handle visual similarity is improperly over-loading LGR.
- Visual similarity issue, if any, should be resolved outside LGR.
- Background reasons
  - "Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels" describes that "While resolving string-confusability issues is beyond the scope of this project, the integration panel will need to ….". This indicates that string-confusability including visual similarity of characters does not necessarily have to be solved in LGR
  - gTLD Applicant Guidebook describes "similarity review will be conducted by an independent String Similarity Panel", which is further augmented by the String Confusion Objection process … as included in the GNSO New gTLD policies. Likewise, string similarity and confusability has been taken into consideration for the IDN ccTLD Fast Track process
  - for any script, visual similarity can only be judged by human intuition which varies with the individuals. And we think solid definition of "visual similarity" must be made only when the definition is universally understandable and precisely definable if it's ever defined in LGR.

Thank You

Update by Korean Generation Panel

Dongman Lee
Member, Korean GP

Agenda

1. Script(s) Covered and where they are used
2. Members of the GP
3. Work achieved to-date 1 (K-LGR v1.0)
4. Work achieved to-date 2 (Coordination between KGP and CGP)
5. Work achieved to-date 3 (Brief History of KGP History)
6. Future Plan and Schedule

Script(s) Covered by K-LGR and Where They Are Used

- K-LGR covers Korean script (= Hangul + Hanja)
- “Korean script” usually means “Hangeul” or “Hangul”. However, in the context of the Korean LGR (K-LGR), Korean script is a union of Hangul (한글) and Hanja (한자).
- Korean language has a long history, more than 2000 years.
- Hangul: invented in 1443.
- Hanja was used before Hangul was invented. Hanja is still used in Rep. of Korea.
- Korean language is mainly used in Rep. of Korea (S. Korea) and Democratic People’s Republic of Korea (North Korea).
  - Also used by Korean people living in China, USA, Japan, Europe, Brazil, Russia, Vietnam, and so on.

Members of the GP

- Technical Experts: Kyongsok KIM (Chair), Dongman LEE
- Linguists: Jeongdo CHOI (Hangul), Sanghyun SHIN (Hanja), Sungduk CHO (Hanja)
- Policy Makers: Youngeum LEE, Youn Jung PARK
- Community: Eunjun JEON, Boknam YUN, Byeongil OH
- Registry: Jinhyun CHO, Minjung PARK, Yunmi CHOI, Ryoung CHAE, Minjee KIM
- Registration Agency: Seong-jin PARK, ChangKi JANG, Myungsoo LEE

Work Achieved To-Date by KGP – 1:

K-LGR v1.0 (2017.12.10.)

- K-LGR v1.0 (2017.12.10.): repertoire and variant groups
  - Hangul: repertoire – 11172 syllables, no variant groups
  - Hanja: repertoire – 4758 characters, 152 variant groups
  - Variant groups composed of Hangul syllables and Hanja chars: 5 (3 Hanja chars: out-of-repertoire variant)

- 4758 Hanja chars in K-LGR v1.0

Source of Hanja Character Set	# chars
1) KS X 1001 (268 comptb. chars excluded)	4620
2) IICORE - K column marked	4744
K-LGR v1.0 (2017.12.10.): Hanja List (Union of 1) and 2))	4758

Work Achieved by KGP – 2:

Public Comments Reviewed

- A summary of public comments
  - Including Hanja in K-LGR repertoire: positive
  - Allowing Hangul-Hanja mixed label: several negative comments, some positive comments
  - Hangul-Hanja variant group: CJK agreement needed
  - Specific details need be corrected/modified
- Examples of issues raised by Mr. Byeon
  - References; quotes; etc.
  - Many Hanja chars allowed for personal names not included in K-LGR
  - Hangul Jamo not included in K-LGR (actually not in MSR-3)
  - More Hangul-Hanja variant groups need be included

Work Achieved by KGP – 2:

Public Comments Reviewed

- Requests by Mr. Byeon for specific details
  - Reviewed and discussed.
  - Mostly accepted in principle and will be reflected in the next version of K-LGR

- Hangul-only labels, Hanja-only labels, Hangul-Hanja mixed labels
  - KGP reconfirmed that there was a general consensus to allow Hangul-only labels and Hanja-only labels;
  - However, KGP has not reached a conclusion as to whether to allow Hangul-Hanja mixed labels.

Plan and Next Steps

- Waiting for the conclusion as to whether to include cross-script (visually identical) variant groups
  - variant groups of Hangul syllables and Hanja characters;
  - variant groups of Kana and Kanji characters
- Hangul-Hanja mixed labels
  - Decide on a final conclusion
- Revision of K-LGR 1.0
  - After the above issues are resolved, K-LGR will be revised and published.

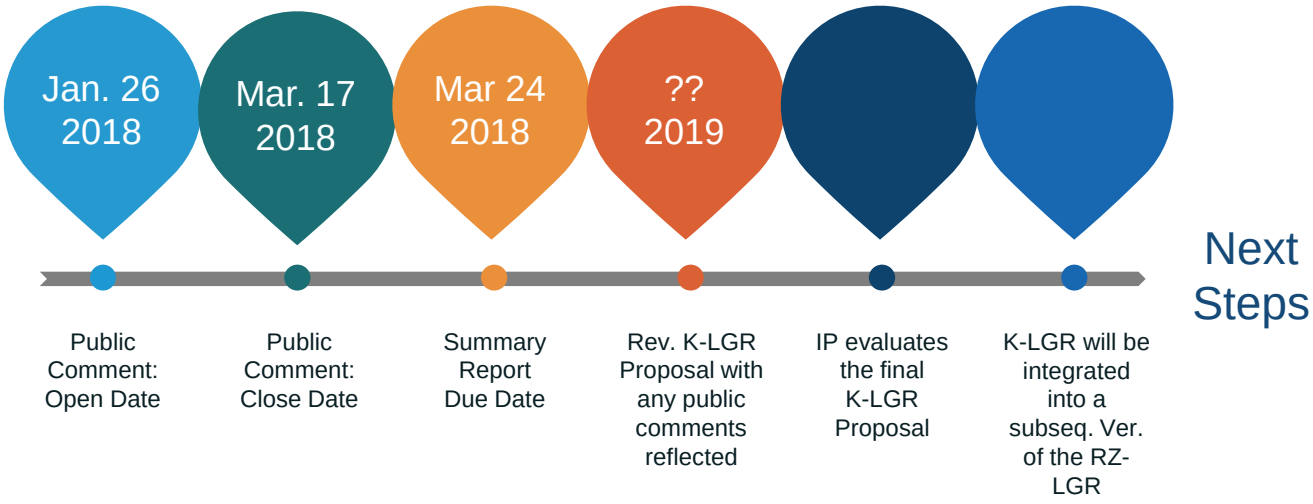
Work Achieved by KGP – 3:

Brief History of KGP Activities

- Dec. 2013: Korean GP (KGP) Organized
- May. 2015: K-LGR v0.1
- Feb. 2016: The Korean community “formally” forms Generation Panel for Developing the Root Zone Label Generation Rules (LGR)
- Dec. 2017: K-LGR v1.0
- Jan. ~ Mar. 2018: public comments for K-LGR v1.0
- Mar. ~ Dec. 2018: public comments for K-LGR v1.0 reviewed for possible reflection in the next version of K-LGR

- 34 KGP meetings
- Several CJK coordination meetings during ICANN meetings 49 ~ 63
- Several CJK coordination meetings in Rep. of Korea, China, and Taiwan.

Future Plans



Engage with ICANN and IDN Program

Thank You and Questions

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

@icann
facebook.com/icannorg
youtube.com/icannnews
flickr.com/icann
linkedin/company/icann
slideshare/icannpresentations
soundcloud/icann