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

#ICANN51

15 October 2014

IDN Root Zone LGR

Sarmad Hussain

IDN Program Senior Manager

#ICANN51

Agenda

- Introduction – Sarmad Hussain
- Need, Limitations and Mechanisms for the Root Zone LGR – Marc Blanchet
- Challenges in Addressing Multiple Languages using Arabic Script– Meikal Mumin
- Coordination between Chinese, Japanese and Korean Scripts – Wang Wei
- Coordination between Neo-Brahmi Scripts – Nishit Jain
- Coordination between Cyrillic, Greek and Latin Scripts – Cary Karp
- Q/A

Types of Coordination

- One script – one GP
  - Arabic
- One script – many GPs
  - Han – Chinese, Korean, Japanese
- Many scripts – one GP
  - Neo-Brahmi scripts
- Many scripts – many GPs
  - Cyrillic, Greek, Latin

Aspects of Coordination

- Need – what work should be undertaken by the GPs
  - Same code points
  - Visually similar code points
  - Similar rules
  - Other?
- Mechanism – how will these GP's interact with each other
  - After individual GP work
  - During individual GP work
  - Before individual GP work

Need, Limitations and Mechanisms for the Root Zone LGR

Presented by: Marc Blanchet

Integration Panel

IDN Root Zone LGR

The Need for LGRs

- It’s not all about variants!

- LGRs define what labels are valid
  - They are needed for automated label validation

- For some scripts, all that is needed is a defined repertoire
  - Each application confined to one repertoire

Root Repertoire

- Collection of single script repertoires
  - Each tagged by script: “und-Cyrl,” “und-Jpan”
  - No cross-repertoire labels
  - No overlap, except “common” code points, Han
- Each script repertoire limited to:
  - Modern, widespread use
  - Everyday use
  - Stable code points

But What About Variants?

- Some scripts require variants
  - Code points that are “the same” to users
- Two types:
  - Those that lead to “blocked” variants
  - Those that lead to “allocatable” variants
- Procedure:
  - Maximize number of blocked variants, and minimize the number of allocatable variants

More on Variants

- Variant mappings will be used to automatically generate all permutations (variant labels)
- Type of variant mapping determines whether:
  - To ***block*** a variant label
(either variant or original can be allocated, not both)
  - To allow ***allocating*** it to the same applicant as original label
- As result of integration, blocked variants can exist across GP repertoires
  - GP coordination will ensure consistent outcome

What, Why and When of WLEs

- Whole Label Evaluation Rules (WLE)
- Why they are needed
  - Prevent labels that cannot be processed/rendered
- When to consider
  - Generally affect “complex scripts”
  - Not intended to enforce “spelling rules”
- Example:
  - Disallow vowel marks where they can’t be rendered: at the start or following other vowel marks, etc.

Limitations

- TLDs are intended for:
  - “*Unambiguous labels with good mnemonic value*” *
- Not intended to capture all facets of a writing system
  - Should focus on modern, everyday use
  - OK not to support some conventions
    - *e.g.*, disallowing apostrophe does not support the ‘s ending for names of businesses, hyphen disallowed in root
  - Some limits necessary to reduce systemic risks

*https://tools.ietf.org/html/draft-iab-dns-zone-codepoint-pples-02

What Should Be Coordinated?

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

Resources

- Considerations for Designing a Label Generation Ruleset for the Root Zone
  - https://community.icann.org/download/attachments/43989034/Considerations-for-LGR-2014-09-23.pdf
- Maximal Starting Repertoire (MSR-1)
  - https://www.icann.org/news/announcement-2-2014-06-20-en
  - https://www.icann.org/en/system/files/files/msr-overview-06jun14-en.pdf
- Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels
  - https://www.icann.org/en/system/files/files/draft-lgr-procedure-20mar13-en.pdf
- Representing Label Generation Rules in XML
  - https://tools.ietf.org/html/draft-davies-idntables
- Requirements for LGR Proposals
  - https://community.icann.org/download/attachments/43989034/Requirements%20for%20LGR%20Proposals.pdf
- Variant Rules
  - https://community.icann.org/download/attachments/43989034/Variant%20Rules.pdf

Challenges in Addressing Multiple Languages using Arabic Script

Meikal Mumin

Arabic Generation Panel

IDN Root Zone LGR

Representing scripts in a world of languages

- abc.def is a Roman/Latin script IDN
- ابت. ثجحis a Arabic script IDN
- But we do not know which languages are used by website of either IDNs
- So International Domain Names (IDNs) have a script as property, but not a language. So what does this mean?
  - It means that IDNs cannot be based on the orthography of one language, such as Arabic language, but that…
  - LGR and related standards must therefore address the entire community of readers and writers of Arabic script
- The problem is that, while we can only represent scripts, we think in terms of language
  - All data is at language level while we have to define LGR at script level
  - There are no institutions representing scripts communities
  - Writing is usually considered as a (reduced) representation of language
- So what is the actual scope of Arabic script LGR?

Scope of the Arabic script LGR

- Arabic script is centered around Africa and the Middle east as a writing system but in the course of time it has expanded across nearly all continents, with established past or present use in
  - the Americas, (Western, Central, Southern, and Eastern) Europe, (nearly all areas of) Asia, Africa (North and South of the Sahara)
  - Only within Africa, there is attested past or present use of Arabic script for the writing of 80+14 African languages apart from Arabic (Mumin 2014)
  - With todays patterns of migrations, continuing proselytization, and population growth, more user communities of Arabic script are manifesting in both the Global South and North
- Accordingly, Arabic script is used not just locally or regionally but globally, albeit to radically different degrees and in entirely different manners, since…
  - for numerous languages, Arabic script is in active competition with other scripts, and…
  - for numerous languages, Arabic script is used only by a part of the language community
  - It is not foreseeable how the situation will evolve in the future and what the impact of IDNs would be on the community
  - To give a more extreme example – Would a language community possibly care if they can register a domain name using the orthography of their language if any reading and writing is only done with pen & paper?

Representing the underrepresented

- Unfortunately, this linguistic diversity is not well represented
    - There is a lack of data on languages and orthographies
    - Particularly languages of low status or socio-economic participation lack representation
    - There is little available on non-western orthographies, while non-standardized orthographies are generally not considered
    - Often much TF-AIDN has to rely on users intuitions from an entirely different part of the script community
    - E.g. during code-point analysis, we frequently lacked data to establish whether a code point is used optionally or obligatorily in a given orthography, which required within the current process

Qualifying and quantifying script use: The EGIDS scale

- Security and stability of DNS and the root zone are highly important, and therefore conservatism is a strong principle surrounding IDNs

  > "Where the Integration Panel was able to establish to its satisfaction that a given code point was assigned a character solely for use in a disused orthography, or for a language in serious decline, the code point has been removed from the MSR."

  Maximal Starting Repertoire — MSR--1 Overview and Rationale, REVISION – June 6, 2014, p. 22

  - MSR dictates that the Expanded Graded Intergenerational Disruption Scale [EGIDS] (Lewis and Simons 2010) is used to categorize the "effective demand" of languages within a given country:
    - The EGIDS consists of 13 levels, ranking languages from the highest representation and role in society, being a National language, to the lowest, extinction
    - "For the MSR the IP used the cut-off between EGIDS level 4 [Educational] and level 5 [Developing]."
- Unfortunately, such representation of language in society is not just accidental but usually a result of historical processes

"Scripts divide languages into cultures, make dialects into new distinct languages, and create new dialects. […] If, as is often said, ‘A language is a dialect with an army and navy’, how much more is it ‘a dialect with a distinct script’!”

(Warren-Rothlin 2014: 264)

People, society, language and the role for IDNs

- Languages and scripts are
  - …evaluated by people (Language attitude)
  - …assigned a status by both societies and scientists (Dialect vs. language)
  - …and regulated by governments (Language policy)
  - and this is reflected also in studies and statistics on languages
    - There have even been historical reports of orthography suppression of Arabic script, where the use of writing systems has been banned and criminalized
- We must be cautious not to strengthen further trends of linguistic discrimination and strive for equal treatment of languages, even where they lack socio-economic participation or political representation
  - TF-AIDN did identify 32 code-points during the analysis, with evidence of use but which cannot be included in LGR because they do not have an EGDIS rating higher than 5

Example #1 – Code point analysis and issues with EGIDS data

- Example Seraiki [ISO 639-3: skr]:
  - Seraiki is a language of Pakistan
  - There are numerous publications in Seraiki, including daily newspapers
  - Within Pakistan, Seraiki has an EGIDS rating of 5 (Written)
  - IP recommends excluding any language with an EGIDS rating lower than 4
- Example Harari [ISO 639-3: har]:
  - Harari is a language of Ethiopia
  - There are significant expatriate communities, which seem to be very active
    - E.g. The Australian Saay Harari Association, which published an orthography description and a virtual keyboard with the assistance of the State Library of Victoria, Australia, in 2009
  - Within Ethiopia, Harari has an EGIDS rating of 6a (Vigorous), while it has not status in Australia
  - Because of the activity of the expatriate community, TF-AIDN assumes an active use of the orthography and would suggest inclusion of relevant code points
  - Unfortunately, this is not possible within the current process stipulated by IP

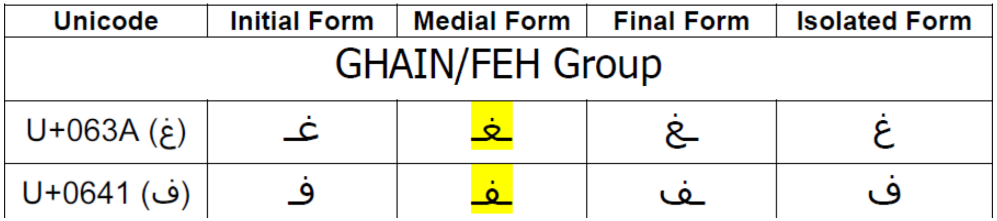
A-priori principles and a-posteriori analysis

- In the case of Arabic script IDNs, ICANN has tasked two groups to work together to develop the Label Generation Rules (LGR)
  - Integration Panel (IP) has developed the "Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels", as well as the Maximal Starting Repertoire (MSR-1)
  - On the basis of the procedure and the MSR-1, the Task Force on Arabic Script IDNs (TF-AIDN), should formulate the LGR, which is then approved by IP
  - Accordingly, rules have been laid out by IP before observation and analysis of data was conducted by TF-AIDN
- Therefore MSR and the LGR development process has been designed before an (ideally data driven) code point analysis could be conducted by script generation panels
  - TF-AIDN noticed this, being the first script generation panel to take up work
  - Accordingly, TF-AIDN did suggest to IP as public comment to MSR-1 that
    - MSR-1 should only be frozen one script at a time
    - after relevant script Generation Panel has been formed and given its feedback on its relevant portion
  - Unfortunately, IP considered this as an effective request for removal of MSR1

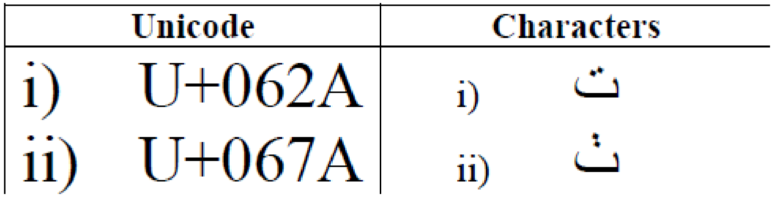
Example #2 - Variants

- Variants are required to balance the usability of IDNs as well as the representation of languages against security and stability of DNS and the root zone

- Arabic Case Study Team Issues Report has published a report, identifying 6 types of variants in Arabic script. Two examples:



Unicode	Initial Form	Medial Form	Final Form	Isolated Form
GHAIN/FEH Group				
U+063A (غ)	غـ	ـغـ	ـغ	غ
U+0641 (ف)	فـ	ـفـ	ـف	ف



Unicode	Characters
i) U+062A	i) ت
ii) U+067A	ii) ٺ

So how can we reasonably argue that this difference in letter shape is *not* confusable by *all* readers and across *all* representations and fonts…

...while this difference is confusable to at least a subset of readers or in a subset of representations and fonts…

- …when there are no empiric scientific tests to support either theory?
- …when there is a systemic bias in representation with even within our group (as 15 out of 29 members are first language speakers of Arabic)?

ەیساک امیرت

شكراً

شکریہ

با سپاس

تشکر

Thank You

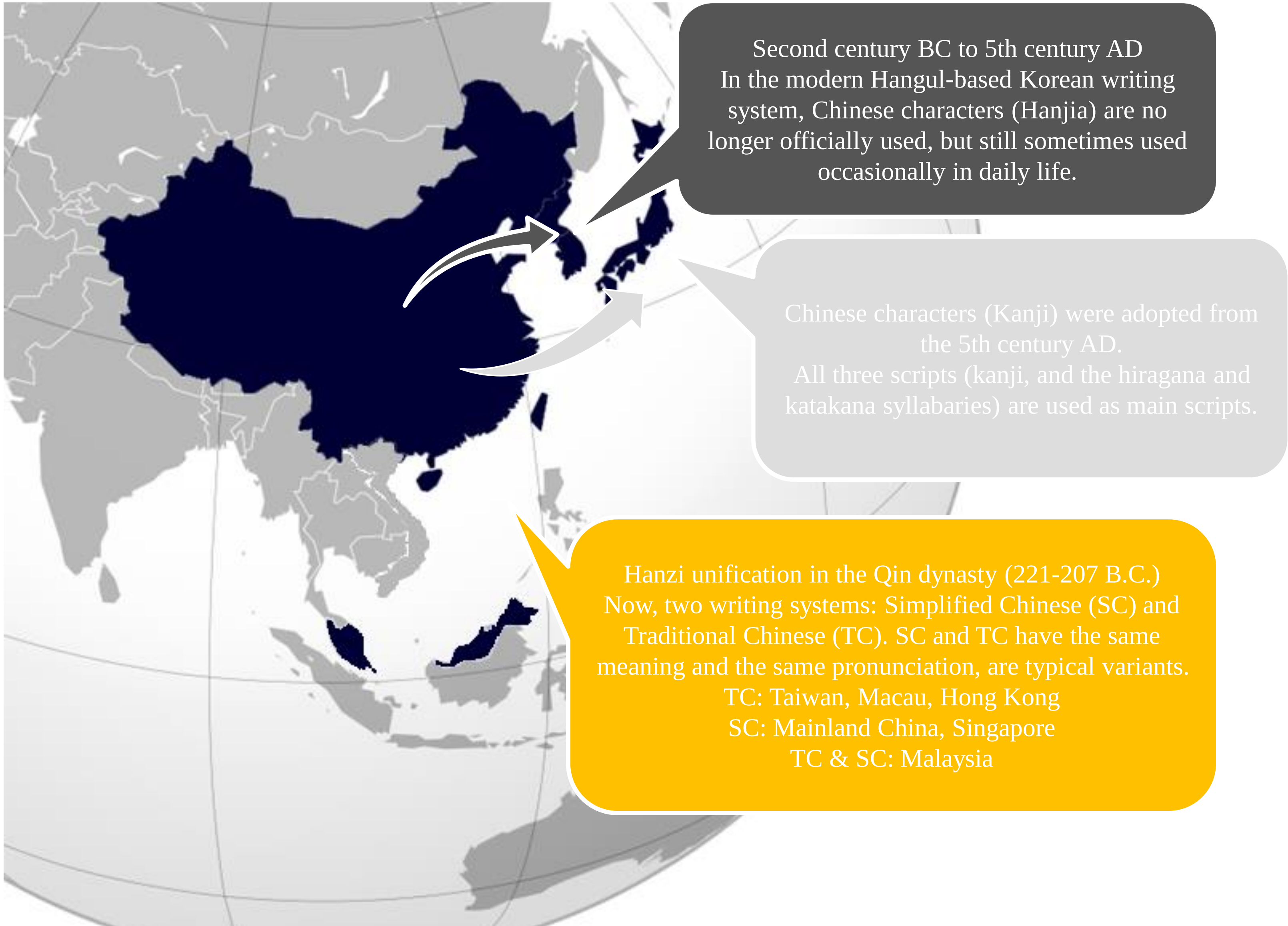
Coordination between Chinese, Japanese and Korean Scripts

Wang Wei

Chinese Generation Panel

IDN Root Zone LGR

The Historical Changes of Chinese Character in East Asia



Second century BC to 5th century AD
In the modern Hangul-based Korean writing system, Chinese characters (Hanjia) are no longer officially used, but still sometimes used occasionally in daily life.

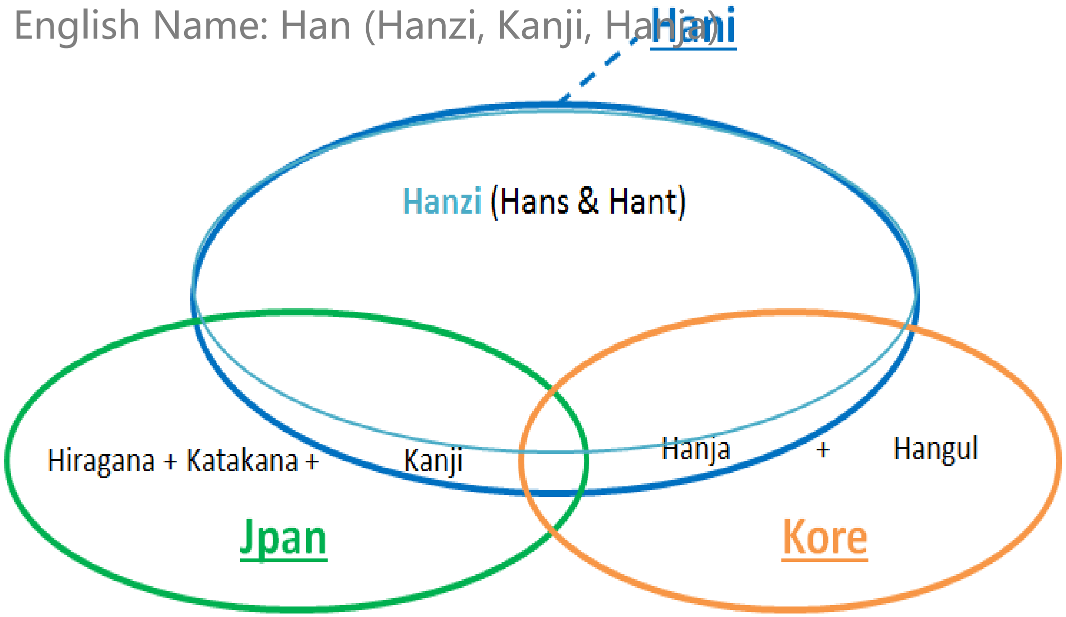
Chinese characters (Kanji) were adopted from the 5th century AD.
All three scripts (kanji, and the hiragana and katakana syllabaries) are used as main scripts.

Hanzi unification in the Qin dynasty (221-207 B.C.)
Now, two writing systems: Simplified Chinese (SC) and Traditional Chinese (TC). SC and TC have the same meaning and the same pronunciation, are typical variants.
TC: Taiwan, Macau, Hong Kong
SC: Mainland China, Singapore
TC & SC: Malaysia

Relationship of Chinese Characters in Three Scripts

In ISO 15924, the script for Chinese characters is mainly defined in this specification:

- ISO 15924 code: Hani
- ISO 15924 no: 500
- English Name: Han (Hanzi, Kanji, Hanja)



SLD/TLD Chinese Character IDN Registration

CDNC Character Table and Registration Rules under RFC 3743/4713

SLD: .CN, .TW, .HK, .SG, .ASIA

TLD: .中国, .台湾, .香港

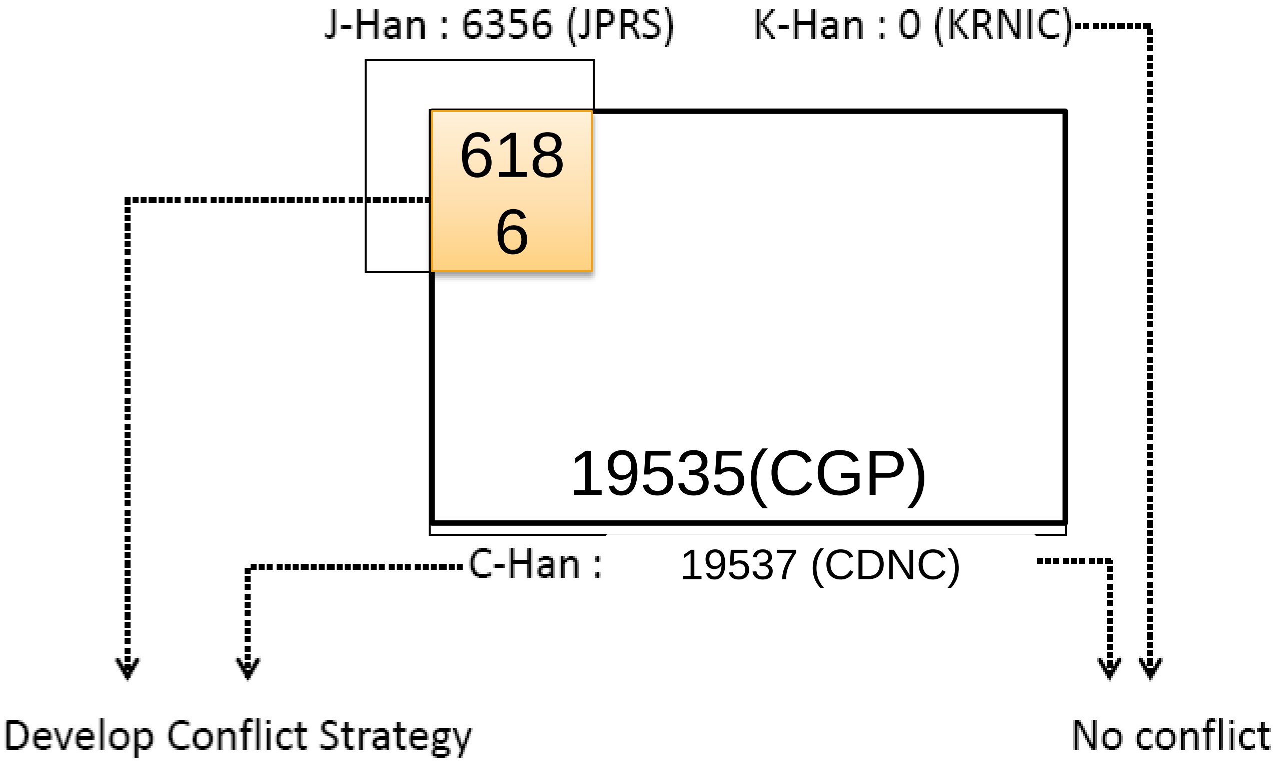
JPRS IDN Registration

SLD: .JP

KISA:

NO Chinese character

registration under .KR



J-Han : 6356 (JPRS)

K-Han : 0 (KRNIC)

618
6

19535(CGP)

C-Han : 19537 (CDNC)

Develop Conflict Strategy

No conflict

So Far

Variant Solutions in Different Scripts

CDNC: RFC 3743 & 4713

- Allocate all Applied-for IDL and Variant IDLs to the same registrant
- Delegate Applied-for IDL, Preferred SC IDL, Preferred TC IDL
- Reserve all the other variant IDLs
- Delegate reserved variant IDLs when requested at a later date

JPRS: No Variant issue

Among Kanji characters, some are in a simplified form (called the “new character form”), derived from the traditional imported form (called the “old character form”).

It is appropriate to distinguish new and old forms as different and independent characters instead of pure variants. This understanding has been reflected in the IANA IDN table developed by the JPRS, in which no variants are identified for Kanji.

KISA: No Variant issue, so far …

Hanja is no longer widely used in the ROK. A law enacted in 2011 orders all ROK official government documents to be written ONLY in Hangul.

KISA stated that its SLD IDN policy does not allow and nor does they have any intention of allowing the use of Hanja in their domestic market.

Coordination Principle

Each CJK panel creates an LGR and each LGR includes a repertoire and variants.

The variant mappings must agree for the same code point for all LGRs.

The variant types may be different (blocked or allocatable), the variant types do not have to agree across LGRs.

The repertoires may be different.

Allocatable

A potential allocation rule says that once the variant label is generated, that variant label may be allocated to the applicant for the original label.

Blocked

A blocking rule says that a particular label must not be allocated to anyone under any circumstances.

Example to Illustrate: Case Study 0
Appendix F of draft-lgr-procedure-20mar13-en.pdf.

For CGP

Code Point	Allocatable Variant	Blocked Variant	Tag
爱 U+7231	愛 U+611B	-	und-hani
愛 U+611B	爱 U+7231	-	und-hani

+

For JGP, probably

Code Point	Allocatable Variant	Blocked Variant	Tag
愛 U+611B	-	-	und-jpab

=

Code Point	Allocatable Variant	Blocked Variant	Tag
爱 U+7231	愛 U+611B	-	und-hani
愛 U+611B	爱 U+7231	-	und-hani
愛 U+611B	-	爱 U+7231	und-jpan

Applying for U+611B using the und-jpan blocks the use of U+7231 in the same location in any label, no matter which tag it is applied under. This is so, even though U+7231 is not a character in Japanese at all and does not appear in the tagged repertoire und-jpan. Because it is not part of that repertoire, it cannot be used in any label applied for with the und-jpan tag.

Progress of CGP, JGP and KGP

CGP:

Formal establishment announcement on 24 September.

(https://www.icann.org/news/announcement-2014-09-24-en)

Draw up initial repertoire and variant type definition in XML format.

Provided some coordination study case for IP and K/J.

JGP: Not seated yet

?

?

KGP: Not seated yet

2014.08.21: KLGP domestic meeting.

2014.08.26: Joint meeting with Han Chuan LEE and other attendees

2014.09.03: CJK people discussion

CGP Repertoire and Variant Type

In 2004, according to RFC 3743 and RFC 4713, CDNC submitted to IANA a unified Chinese Character Set (19520 characters) for domain name registration, building up mapping relationships between any given simplified character, its traditional character(s) and its variant(s).
In 2012, CDNC added 17 more Chinese characters as requested by Hongkong community, increasing the set number to 19537. But only 15 of those 17 characters are included in MSR-1.

- Thus CGP takes **the intersection of MSR-1 and the latest version of CDNC character set**, amounting to **19535** characters, excluding Latin Hyphen, digits and letters.

- Following CDNC registration rule and RFC 3743 & 4713, CGP take the second column (the preferred variants) as “allocatable,” while the rest of the variants as “blocked.”

```
<char cp="575D" tag="sc:Hani">
<var cp="575D" type="simp" comment="identity" />
<var cp="57BB" type="block" />
<var cp="58E9" type="trad" />
</char>
<char cp="57BB" tag="sc:Hani">
<var cp="575D" type="simp" />
<var cp="57BB" type="block" comment="identity" />
<var cp="58E9" type="trad" />
</char>
<char cp="58E9" tag="sc:Hani">
<var cp="575D" type="simp" />
<var cp="57BB" type="block" />
<var cp="58E9" type="trad" comment="identity" />
</char>
```

Code Point	Allocatable Variant	Blocked Variant	Tag
坝 (575D)	壩 (58E9)		und-hani
坝 (575D)		垻 (57BB)	und-hani
垻 (57BB)	坝 (575D)		und-hani
垻 (57BB)	壩 (58E9)		und-hani
壩 (58E9)	坝 (575D)		und-hani
壩 (58E9)		垻 (57BB)	und-hani

CGP’s Perspective for Variant Mapping Coordination

- CGP is aware that the coordination can not be achieved by one party.
- CGP is tremendously open to make an unified variant mapping table working together with JGP and KGP.
- CGP is ready to modify the initial repertoire and variant type annotation according to the coordination result, and if necessary, to delete some code points to avoid complicated conflicts.
- Those UNIQUE Chinese character codes in JGP and KGP are NOT to be added into CGP repertoire.

Case Study 1

All code points are included in CGP initial repertoire and regarded as variants of each other.

The mapping relationship in RFC 3743 format is as follows:

- 一4E00 (0); 一4E00(86),一4E00(886); 一(0),壱(0),壹(0),弌(0);
- 壱58F1 (0); 壹58F9(86),壹58F9(886); 一(0),壱(0),壹(0),弌(0);
- 壹58F9 (0); 壹58F9(86),壹58F9(886); 一(0),壱(0),壹(0),弌(0);
- 弌5F0C (0); 一4E00(86),一4E00(886); 一(0),壱(0),壹(0),弌(0);

Meanwhile, all code points are included in JPRS IDN table as well.

(http://www.iana.org/domains/idn-tables/tables/jp_ja-jp_1.2.html)

There is no mapping relationship among them.

一	4E00(2,3);4E00(2,3);	# 16-76, CJK UNIFIED IDEOGRAPH-4E00
壱	58F1(2,3);58F1(2,3);	# 16-77, CJK UNIFIED IDEOGRAPH-58F1
壹	58F9(2,3);58F9(2,3);	# 52-69, CJK UNIFIED IDEOGRAPH-58F9
弌	5F0C(2,3);5F0C(2,3);	# 48-01, CJK UNIFIED IDEOGRAPH-5F0C

Case Study 1

Code Point	Allocatable Variant	Blocked Variant	Tag
一 (U+4E00)	-	壱 (U+58F1)	und-hani
一 (U+4E00)	-	壹 (U+58F9)	und-hani
一 (U+4E00)	-	弌 (U+5F0C)	und-hani
壹 (U+58F9)	-	一 (U+4E00)	und-hani
壹 (U+58F9)	-	壱 (U+58F1)	und-hani
壹 (U+58F9)	-	弌 (U+5F0C)	und-hani
弌 (U+5F0C)	一(U+4E00)	-	und-hani
弌 (U+5F0C)	-	壹 (U+58F9)	und-hani
弌 (U+5F0C)	-	壱 (U+58F1)	und-hani
壱 (U+58F1)	壹(U+58F9)	-	und-hani
壱 (U+58F1)	-	一 (U+4E00)	und-hani
壱 (U+58F1)	-	弌 (U+5F0C)	und-hani
一 (U+4E00)	-	-	und-jpan
壹 (U+58F9)	-	-	und-jpan
弌 (U+5F0C)	-	-	und-jpan
壱 (U+58F1)	-	-	und-jpan

Case Study 2

The code point and its variant(s) exist **separately** in CGP and JGP

- 刊 (U+520A) # in CGP and JGP
- 刋 (U+520B) # in CGP and JGP
- 栞 (U+681E) # only in JGP

In CGP repertoire, the mapping is:

- 刊520A (0);刊520A(86),刊520A(886);刊(0),刋(0);
- 刋520B (0);刊520A(86),刊520A(886);刊(0),刋(0);

In JPRS table，code points are:

- 刊 520A(2,3);520A(2,3);
- 刋 520B(2,3);520B(2,3);
- 栞 681E(2,3);681E(2,3);

Case Study 2

Though 栞(U+681E) is not included in CGP repertoire, but it is regarded as the variant of 刊 (U+52-A) and 刋(U+520B) in ancient Chinese literature and some local areas**.**

CGP would like to extend the CGP repertoire by adding 栞(U+681E) and build up the variant relationship.

Code Point	Allocatable Variant	Blocked Variant	Tag
刊(U+520A)	-	刋(U+520B)	und-hani
刊(U+520A)	-	栞(U+681E)	und-hani
刋(U+520B)	刊(U+520A)	-	und-hani
刋(U+520B)	-	栞(U+681E)	und-hani
栞(U+681E)	刊(U+520A)	-	und-hani
栞(U+681E)	-	刋(U+520B)	und-hani
刊(U+520A)	-	-	und-jpan
刋(U+520B)	-	-	und-jpan
栞(U+681E)	-	-	und-jpan

Case Study 3

The code point ONLY exists in JPRS table:

- 辻(U+8FBB)

‘辻’ does NOT exist in CGP now and traditionally, it is regarded as a Japanese UNIQUE character code.

If CGP linguistic experts keep the viewpoint that ‘辻’ is not associated any code point in CGP repertoire, CGP will not add this code point into CGP repertoire：

Code Point	Allocatable Variant	Blocked Variant	Tag
辻(U+8FBB)	-	-	und-jpan

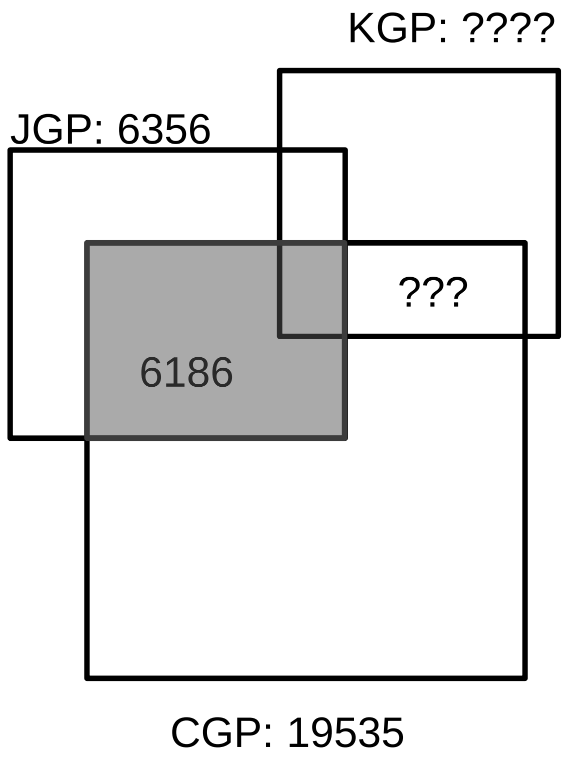
Expectation for JGP and KGP

- Generate the repertoire and variant type annotation ASAP
  - JGP: Kanji repertoire and variant type annotation
  - KGP: Allow Hanjia? >> Hanjia repertoire and variant type annotation

- Work together on the unified variant mapping table for the overlapped code points
  - Case 0: jpan or kore tagged code point block hani variant
  - Case 1: NO change to any variant type annotation
  - Case 2: jpan or kore tagged code point added into hani variant
  - Case 3: jpan or kore UNIQUE code points
  - Case 4: …

- Revise each panel's repertoire and variant type annotation and cross-check the consistency and potential conflicts.

- Generate each panel's Whole Label Generation Rule and cross-check the consistency and potential conflicts.



Challenges…

- Postponed work plan
  - Synchronization between C, J and K
  - Extension from 31 Dec 2014 to 2015
- Repertoire Modification
  - Negotiation among three panels’ linguistic experts
  - Code points extension or reduction
  - Variant type annotation changes
- Whole Label Generation Rule Set
  - Each panel SHOULD be aware of PROs and CONs of the language tag based solution
  - Focus on the techniques and best-practice

Thanks

Q&A

Coordination between Neo-Brahmi Scripts

Nishit Jain

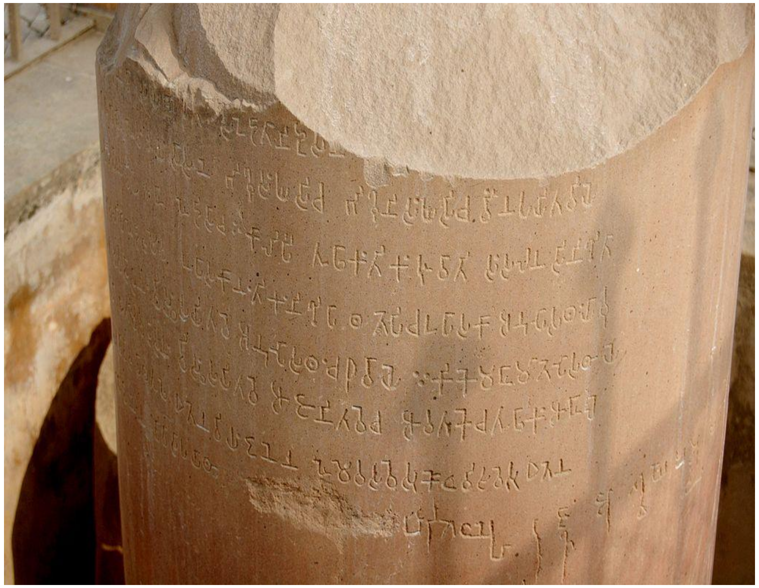
Neo-Brahmi Generation Panel

IDN Root Zone LGR

Neo-Brahmi Generation Panel

What is Brahmi?

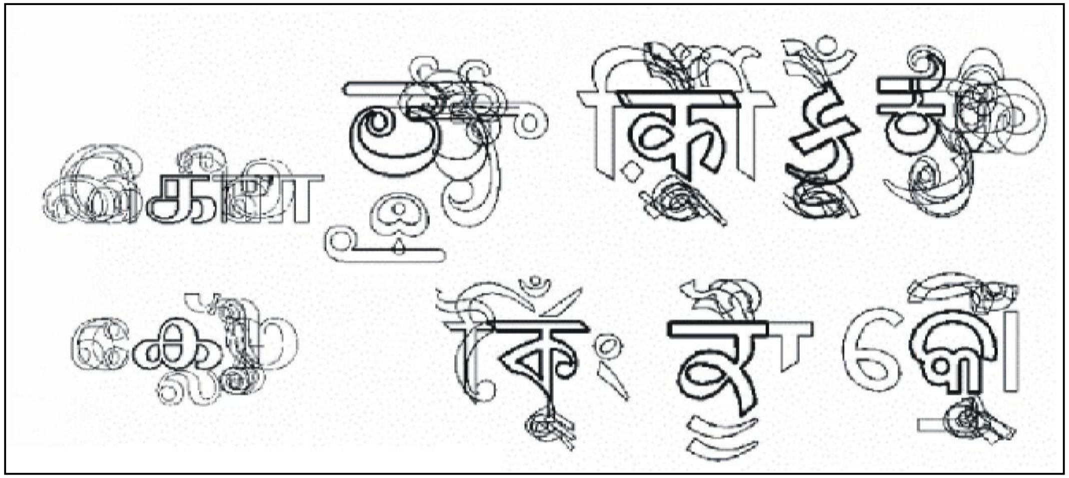
- An ancient script
- Most of the modern scripts in Indian subcontinent have been derived from Brahmi.
  - Geographically the scripts being used in Central Asia, South Asia and South-East Asia
- These scripts are used by multiple language families: Largely by Indo-Aryan and Dravidian



Brahmi script engraved on Ashoka Pillar in 3rd century BCE
Source:
http://en.wikipedia.org/wiki/Brahmi_script

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.



Why Neo-Brahmi?

- 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.

Previous Similar Work

- For IDN version of “.in” ccTLD, (.bharat) equivalent in 22 Official Indian Languages, similar exercise had been carried out

- Following things were finalized for each language
  - Permissible set of code points
  - Visually similar variant strings
  - Complex whole label evaluation rules

- Recently .भारत ccTLD has been launched in Devanagari script covering Hindi, Marathi, Konkani, Boro, Dogri, Maithili, Nepali and Sindhi.

Revisiting the Rules in Context of LGR Framework

- LGR work is different in following contexts
  - Wider stakeholder group
  - Overarching principles in the LGR procedure
    - Especially *Simplicity* and *Predictability* principles
- This revision, however, would not change
  - The need for the well-formedness of the label in terms of Akshar formalism

Neo Brahmi GP - Current Status

- Currently the group is 10 members
  - Mixed bag of expertise like linguistic, Unicode

Udaya Narayana Singh	Raiomond Doctor
Mahesh D. Kulkarni	Anupam Agrawal
Akshat S. Joshi	Abhijit Dutta
N. Deiva Sundaram	Neha Gupta
Nishit Jain	Prabhakar Pandey

  - We are in process of getting more members on-board

Neo Brahmi GP – Outreach Efforts

- Conducted a workshop in AprIGF-2014 for awareness and call for participation in LGR procedure.
    - Topic: "*Bringing diverse linguistic communities together for a unified IDN ruleset*"
    - The panel discussion touched upon the various aspects of creation of the LGR for the Neo-Brahmi scripts
    - http://2014.rigf.asia/agenda/workshop-proposals/workshop-proposal-13/
- Participation and presentation in ICANN 49 public meeting at Singapore
- Participation and presentation in ICANN 50 public meeting at London

Reaching out to the community for wider participation

Cross-Script Similarities

DEVANĀGARĪ SCRIPT	COGNATE SCRIPT	CODEPOINT IN COGNATE SCRIPT
घ U+0918	Gujarati	ઘ U+0A98
उ U+0909	Gurmukhi	ਤ U+0A24
र U+0930	Gujarati	ર U+0AAE

- Code point similarity across scripts
- Cases where Devanagari-Gujarati and Devanagari-Gurumukhi strings look similar.

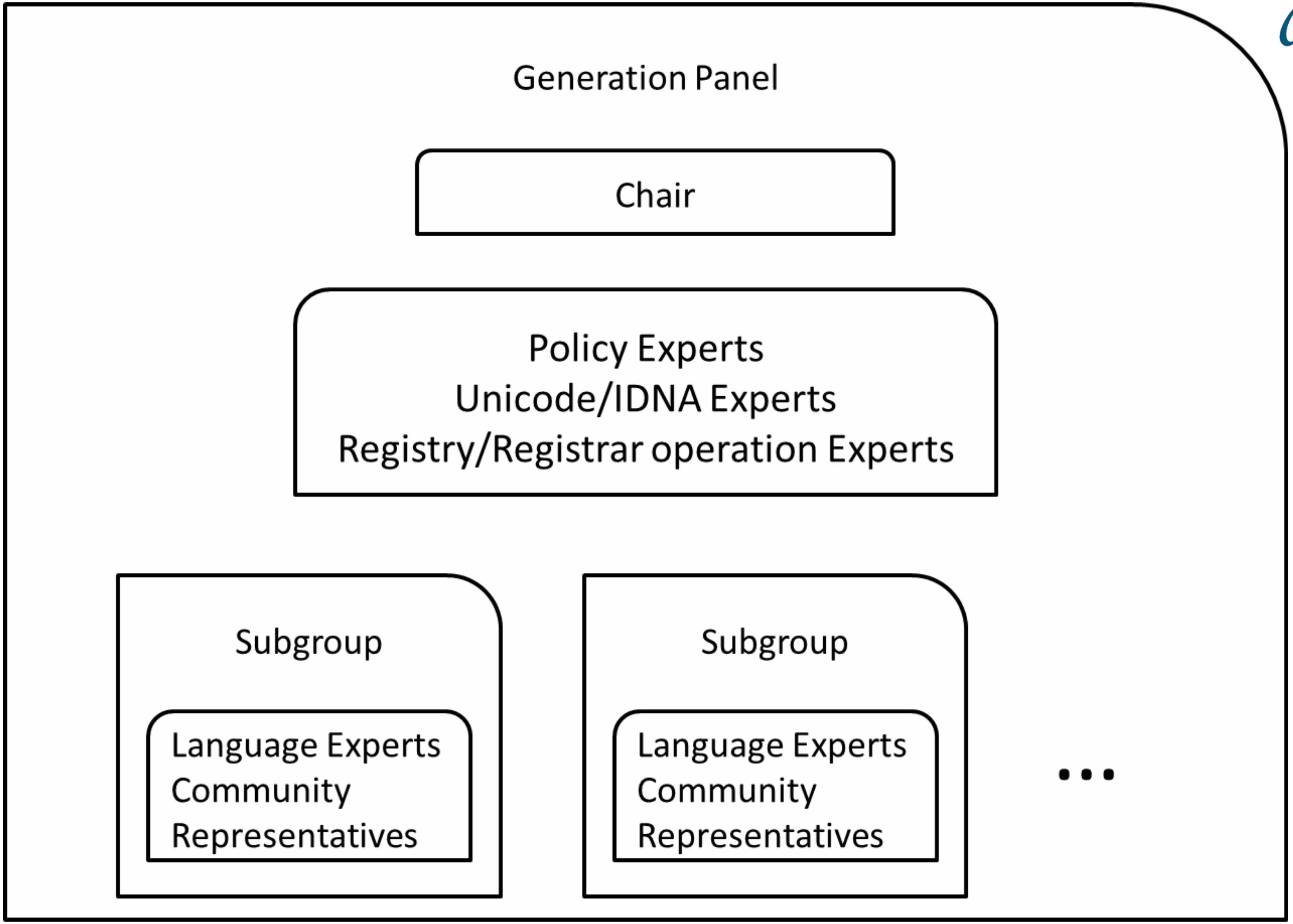
घर ઘર

U+0918 U+0930 U+0A98 U+0AAE

घटी ਬਟੀ

0918 091F 0940 0A2C 0A1F 0A40

Neo Brahmi GP – Approach



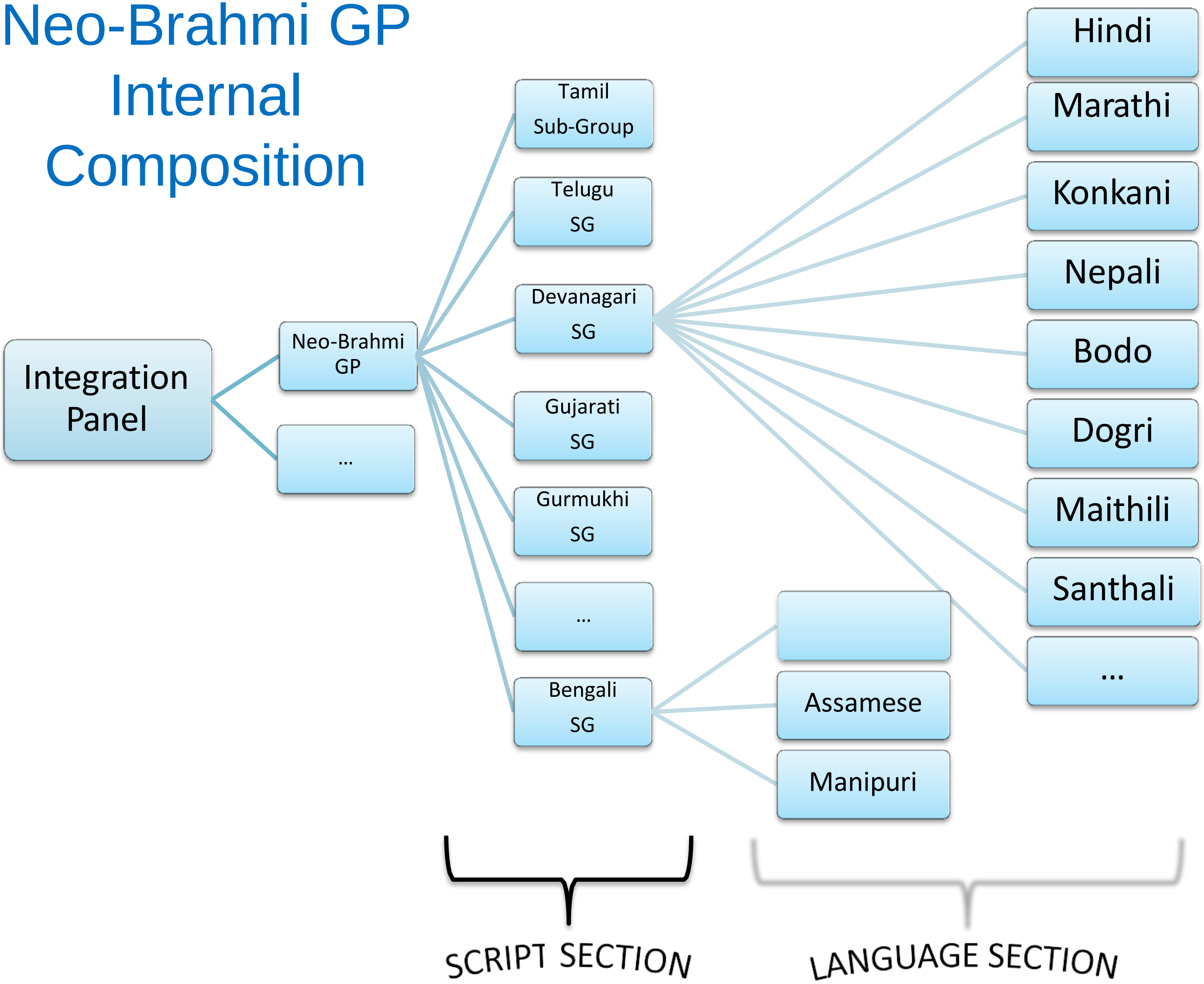
Neo Brahmi GP – Approach

- There are cases of:
  - One script, one language
  - One script, multiple languages

- Multiple sub-groups may exist to ensure proper representation of each language

- Each sub-group ideally would comprise of
  - Language expert(s)
  - Community representative(s)

Neo-Brahmi GP Internal Composition



Integration Panel

Neo-Brahmi GP

…

Tamil Sub-Group

Telugu SG

Devanagari SG

Gujarati SG

Gurmukhi SG

…

Bengali SG

Hindi

Marathi

Konkani

Nepali

Bodo

Dogri

Maithili

Santhali

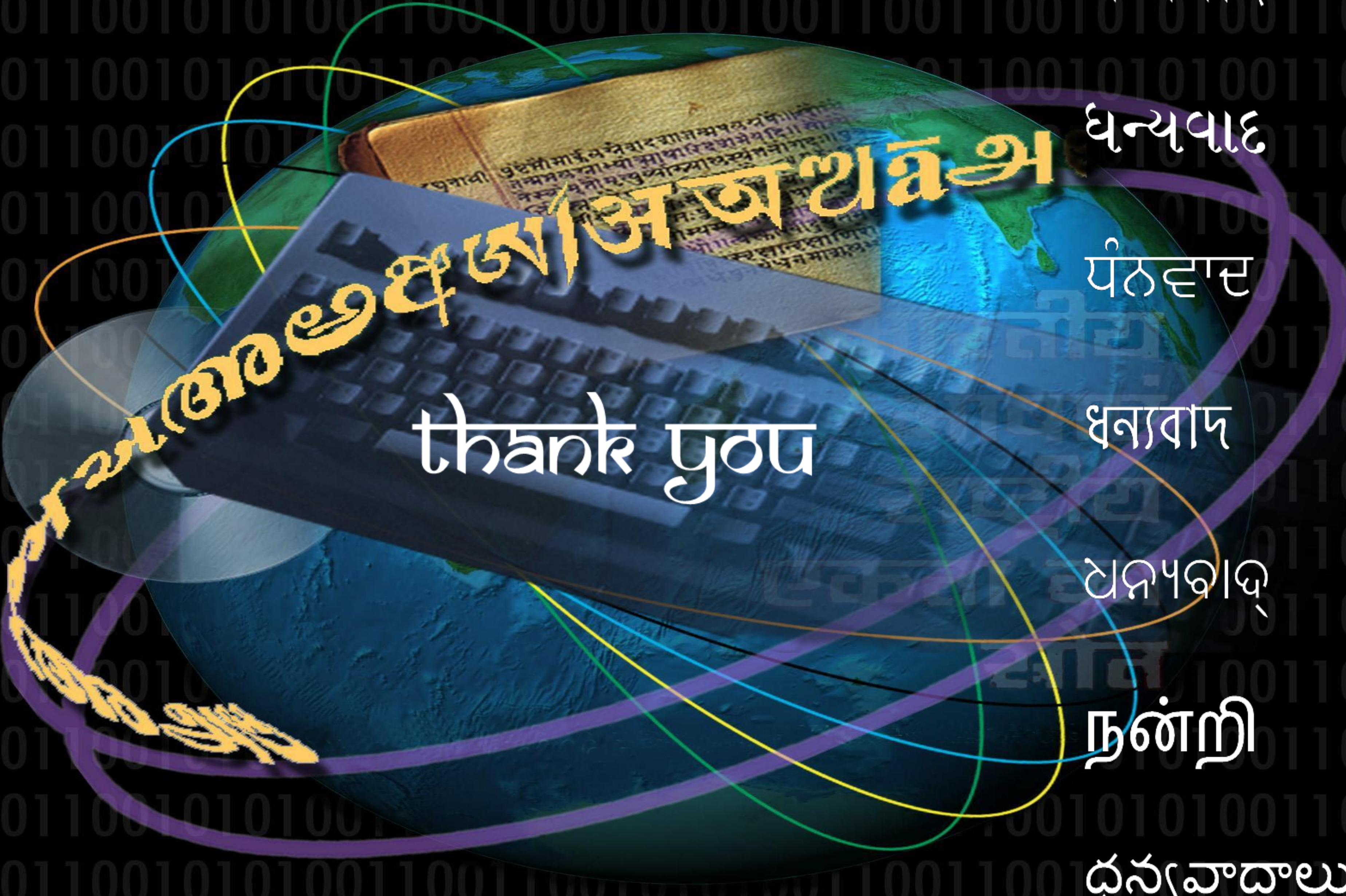
…

Assamese

Manipuri

SCRIPT SECTION

LANGUAGE SECTION



thank you

धन्यवाद

ધન્યવાદ

ਧੰਨਵਾਦ

ধন্যবাদ

ଧନ୍ୟବାଦ୍

நன்றி

ధన్యవాదాలు

Coordination between Cyrillic, Greek and Latin Scripts

Cary Karp
Latin Generation Panel
IDN Root Zone LGR

- **Apsyeoxic** — two words that appear to be spelled identically but are actually sequences of characters from different scripts are said to be **apsyeoxic** /æpsiˈaːksɪk/. This term is derived from the graphic similarity between the string of Roman letters apsyeoxic and the visually confusable string of Cyrillic letters apsyeoxic

  ○ http://dictionary.sensagent.com/

- Culling Cyrillic and Latin code points from the MSR which are commonly represented with congruent glyphs:
  - Latin
    - aäæcçdeëəhiïoöpsxyÿʒ
  - Cyrillic
    - аӓӕсҫԁеёәһіїоӧрѕхуӱӡ

- Adding Greek and admitting closely similar, but not identical glyphs:
  - Cyrillic
    - а ҫїко р
  - Greek
    - αβγςϊκοόρν
  - Latin
    - ɑßɣçï oópv
- The extent of the problem crossing all three scripts does not appear particularly great

- Stepping away from both IDNA and the MSR and considering uppercase:
    - Cyrillic
        - АГВЕНІКМ ОПРТФХ
    - Greek
        - ΑΓΒΕΗΙΚΜΝΟΠΡΤΦΧΥΖ
    - Latin
        - A BEHIKMNO PTɸXYZ
- IDNA expects issues relating to case to be resolved before the protocol is invoked
- This does not mean that such issues are irrelevant

- This does mean that if the LGR panels are to address cross-script issues, they may also need to deal with collateral details that lie outside the current scope of the initiative

Thank You

Engage with ICANN on Web & Social Media

twitter.com/icann

facebook.com/icannorg

linkedin.com/company/icann

youtube.com/user/ICANNnews

gplus.to/icann

weibo.com/icannorg

flickr.com/photos/icann

icann.org

ICANN