ICANN | 54
Dublin
18-22 OCTOBER 2015

IDN Root Zone LGR Workshop

Sarmad Hussain | ICANN54 | 21 October 2015

Agenda

- ⊙ Root Zone LGR   – Asmus Freytag, IP

- ⊙ Community Updates

  - ⊙ Chinese GP update   – Wang Wei

  - ⊙ Japanese GP update   – Hiro Hotta

  - ⊙ Arabic Script GP update – Nabil Benamar

  - ⊙ Armenian GP update   – Igor Mkrtumyan

  - ⊙ Cyrillic GP update   – Yuriy Kargapolov

- ⊙ Q/A

Root Zone LGR

Asmus Freytag

Integration Panel

Agenda

- ⊙ Requirements for LGR Proposals
  - • Format
  - • Common issues

- ⊙ Maximal Starting Repertoire Status

- ⊙ LGR Review
  - • Review process
  - • Integration

Requirements for LGR Proposals
What the IP is looking for
in an LGR submission

Required Elements

- The first LGR proposals have been submitted for public comment. The IP is reviewing these to help identify any issues before final submission.

- Some common issues are seen in draft proposals.

- Required elements of an LGR proposal are spelled out in "Requirements for LGR Proposals"
  - https://www.icann.org/en/system/files/files/Requirements-for-LGR-Proposals-20150424.pdf

- Required documents:
  - Main Document
  - XML formatted tables
  - Label files (test cases: valid and invalid)

Format of Submission

- ◉ Use LGR proposal template for a consistent layout and appearance of the main document
  - https://community.icann.org/download/attachments/43989034/LGR-Proposal-Template.docx

- ◉ Common XML file issues
  - Must contain default WLE rules and actions from MSR
  - UI tool for LGR editing and error checking,
    see workshop presentation
  - Issues with references, see next slide

- ◉ IP is happy to review early drafts and help with format

References cited in the LGR

Based on the experience with the first few LGR drafts, the Integration Panel would like to highlight these requirements:

⦿ Each code point should have an associated reference
- The reference should establish that the code point is in current use and appropriate for TLD labels
  - o For some scripts that can be challenging

⦿ In the XML file, use numeric reference IDs starting at 100
- Make sure that the main document also uses these numbers
- For all other references In the main document use standard "[reference]" –style alphabetic citation

⦿ Delete unused references, make references consistent across files

Other issues

- Provide a citation for the [MSR] version the LGR is based on

- Versioning:
  - The XML file's <version> element is the LGR version
  - The document may also give a document version

- More common XML file issues:
  - Should defer to description document for most details
  - Except those needed to be able to read the XML file at hand
    - Link to description document (must get from ICANN staff)
    - Cite correct version of description document
  - Must preserve default rules and actions, but otherwise
    - Delete unused rules, classes, actions, etc.

MSR-2

Status and Plans for the Maximal Starting Repertoire

Maximal Starting Repertoire (MSR)

◉ [MSR-2 is still the current version](#)

◉ Cannot add new code points from latest versions of Unicode because IDNA2008 process has stalled
  - Latest Unicode version is 8.0
    - ○ IDNA2008 is stuck at 6.3

◉ No changes are planned within the Unicode 6.3 repertoire
  - No pending requests for code points or scripts
  - IP has no concrete plans for additional scripts

◉ As of the present time, MSR-3 is on hold until IDNA2008 restarts

IP Approach for LGR Review

"Failure is not an option"

ICANN

IP Approach to LGR Review

⊙ IP can only give up/down vote, cannot modify LGRs to fix them

⊙ Structures LGR review process to reduce chance of rejection:

- Early discussion with GPs to discover any issues before LGR proposal goes to public comment

- Engage in public comments to resolve any outstanding issues

- Reduce the chance of "surprises" during review for integration, increase chance of accepting LGRs

⊙ Final review and integration of related scripts requires their LGRs be considered together, this may affect the timeline

Integration Process

- Integration
    - Review and final acceptance
    - Integration as defined in procedure
    - No technical changes, unless permitted for integration
    - May reformat or restate XML

- In XML format, the "integrated" LGR is a collection of per-script files
    - given an input label AND a ISO 15924 script code
        - use the LGR corresponding to the script code to process the input label

- Scope of any public comment after integration limited to:
    - Verification of integration
    - Editorial issues
    - No technical changes

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
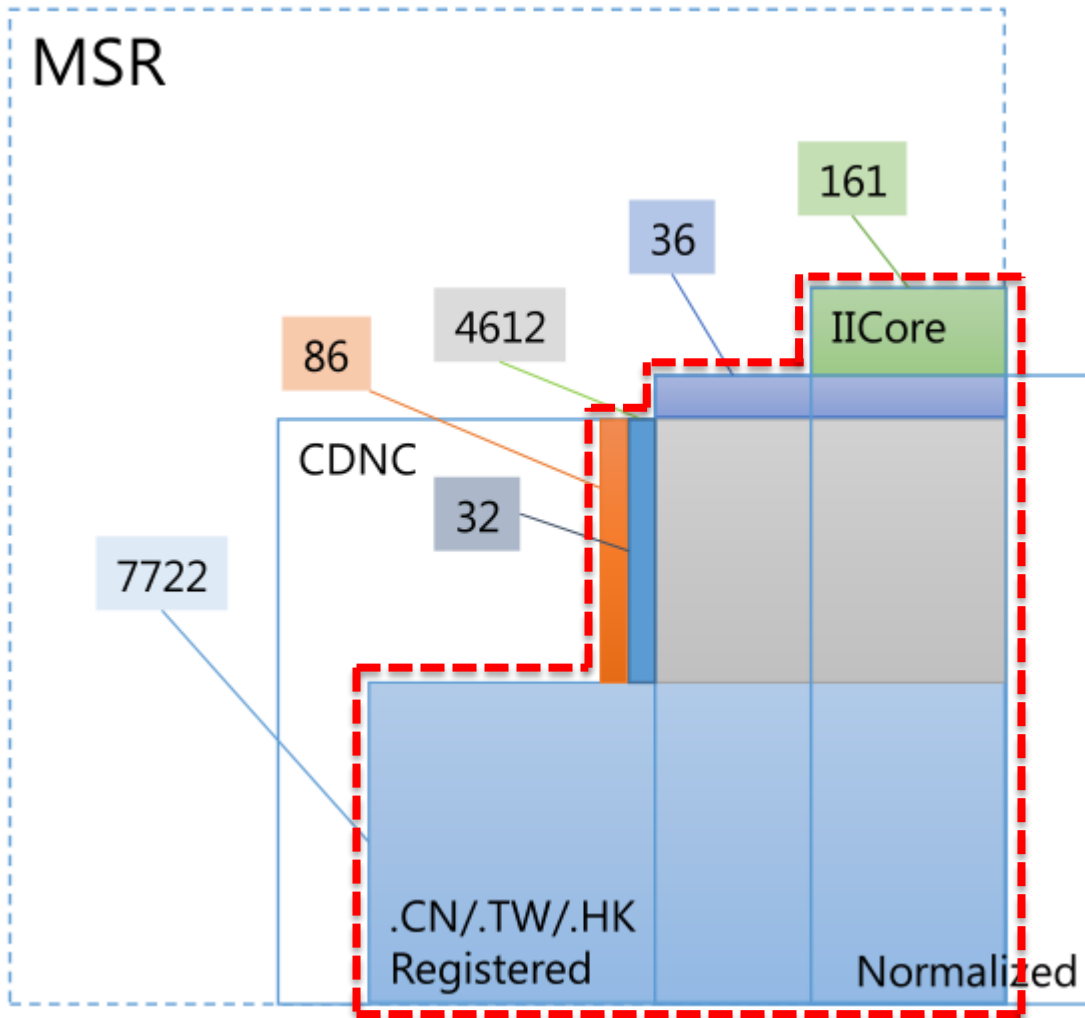
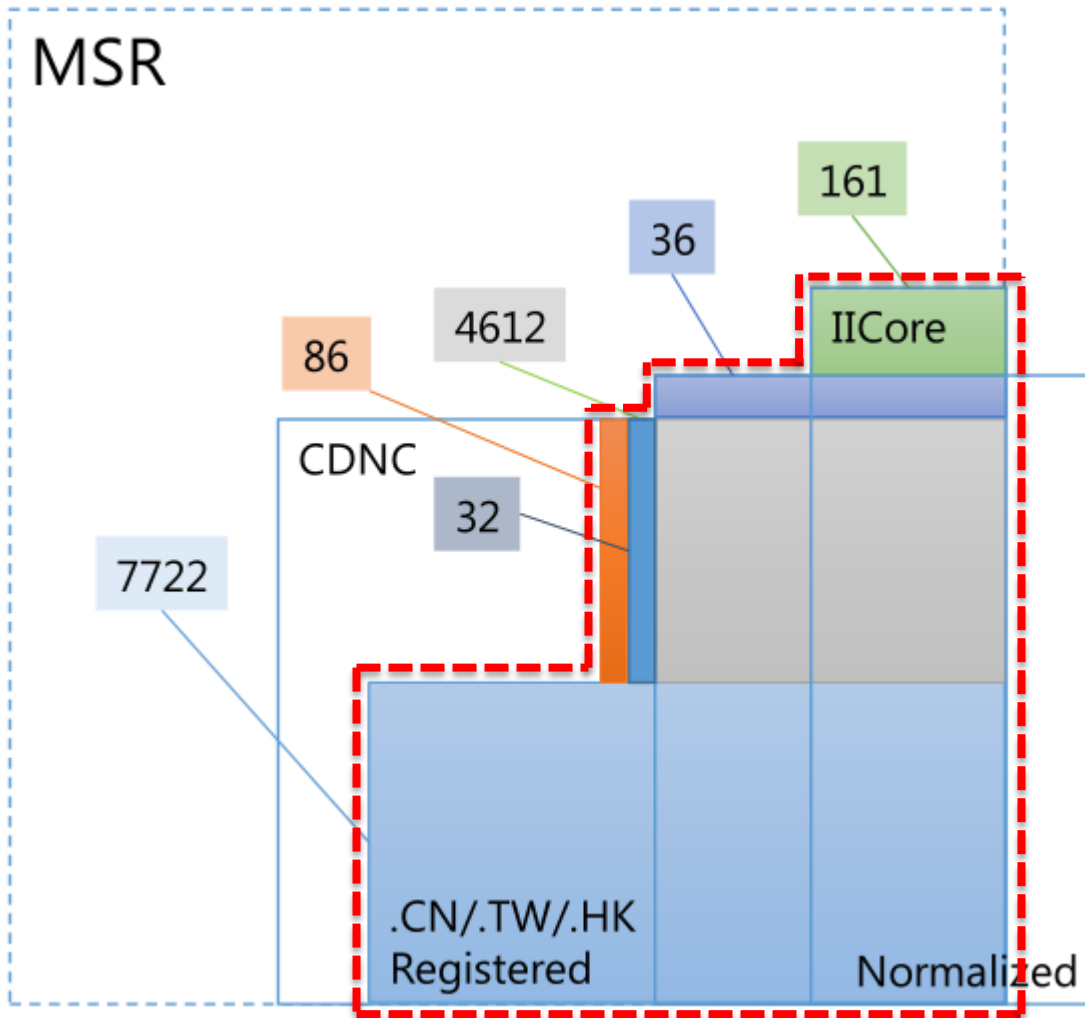# Chinese Generation Panel
Wei WANG and Kenny HUANG

ICANN

Minimum Shared Set

CGP think a reduction of CDNC table (19559 chars, mainly reflects Chinese language community's concern)is not enough to support a coordination across three language environment.

As an alternative, MSS is proposed to work as the basis of coordination between C, J and K.
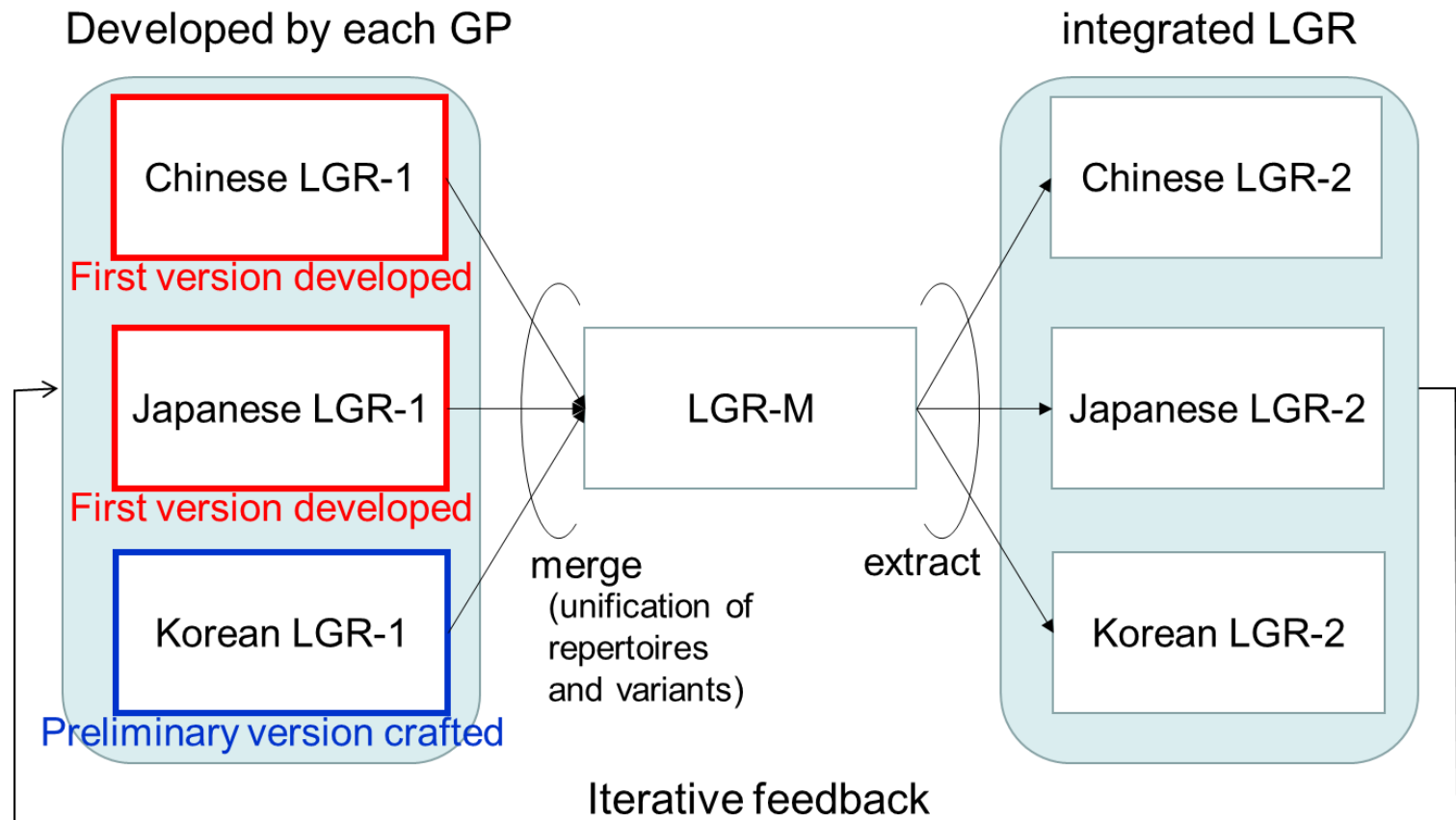
Minimum Shared Set

- Historically registered chars under .CN/.TW/.HK
  - 7722
- Normalized Table by government
  - 4612+36
- IICore by Asia countries
  - 161
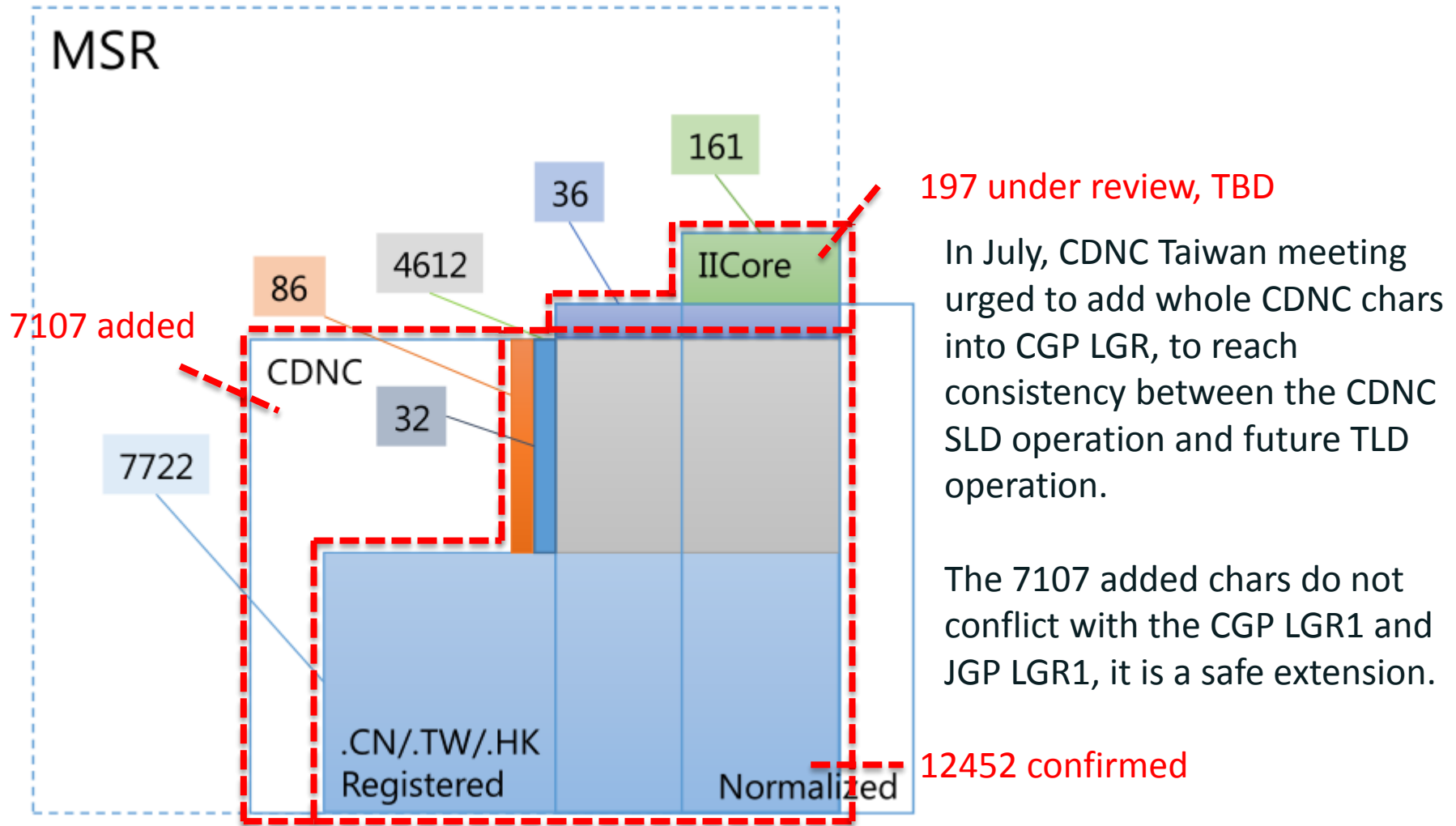- variants to normalized table and IICore
  - 32+86
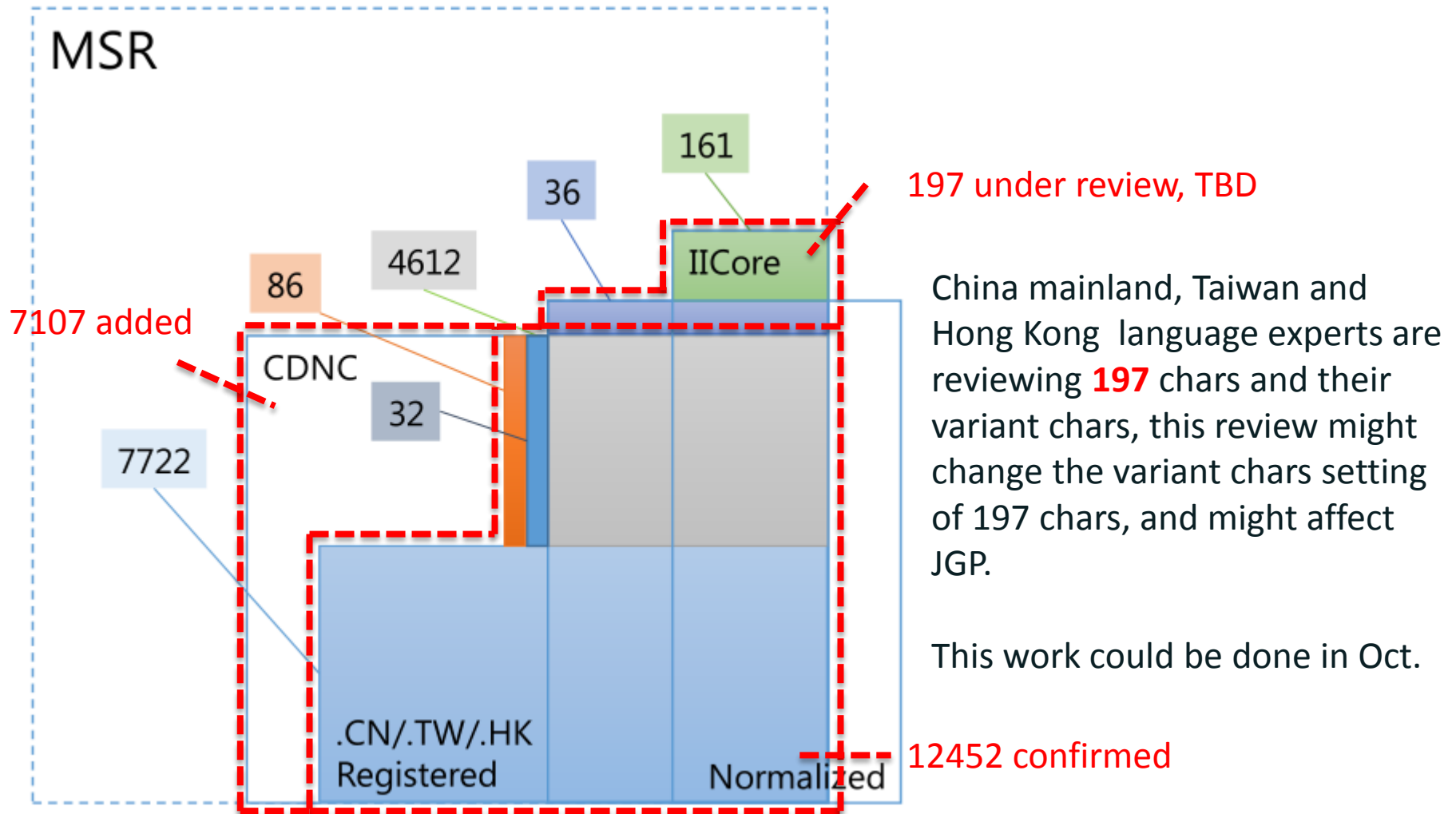
12649 chars

MSS >> CGP LGR-1

In June, CGP and JGP input C LGR-1 (MSS) and J LGR-1, **got the same output as LGR-2**.

MSS to CDNC+(CGP LGR alpha)

MSR

7107 added

86
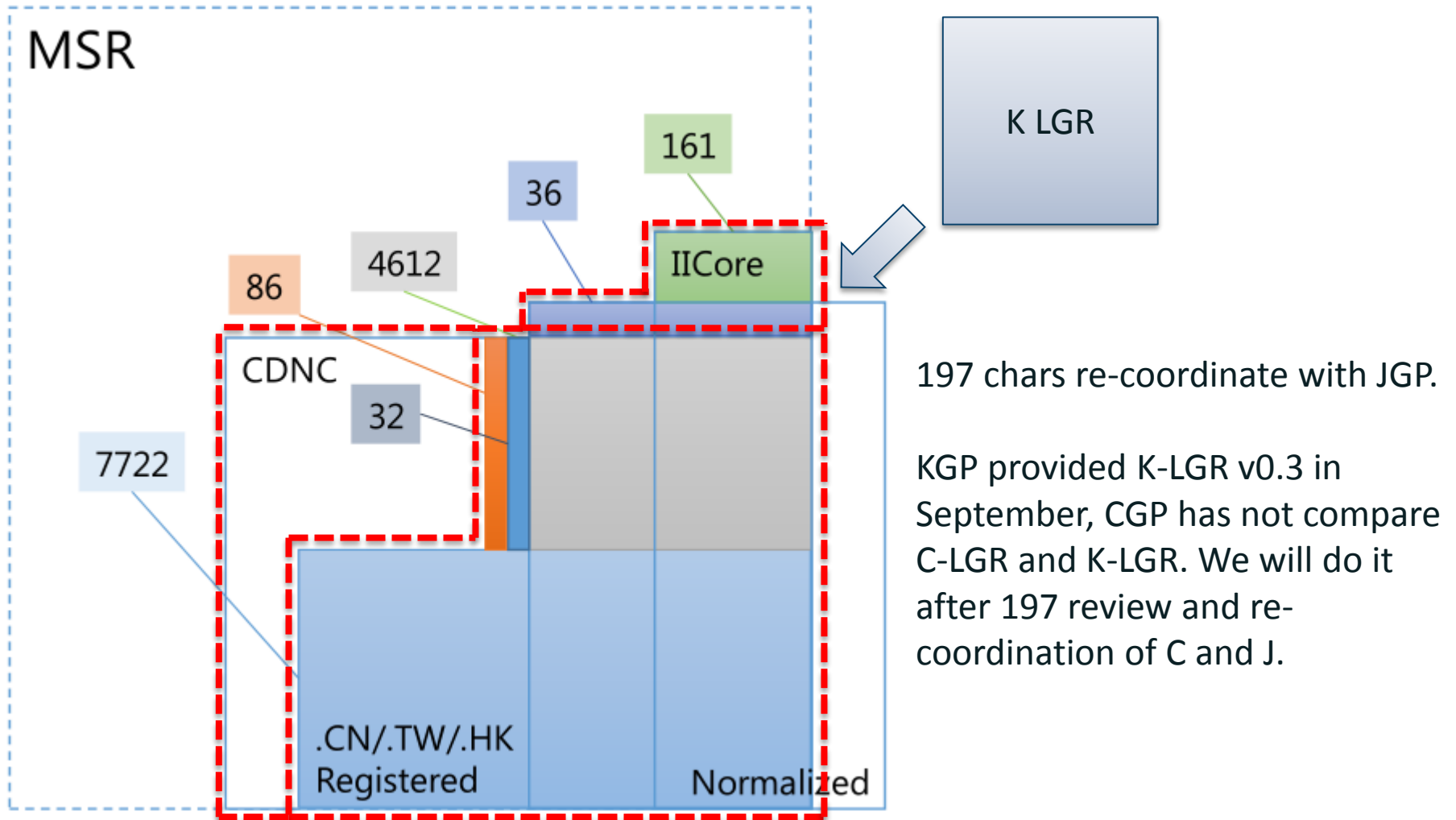
4612

36

161

IICore

32

CDNC

7722

.CN/.TW/.HK Registered

Normalized

197 under review, TBD

In July, CDNC Taiwan meeting urged to add whole CDNC chars into CGP LGR, to reach consistency between the CDNC SLD operation and future TLD operation.

The 7107 added chars do not conflict with the CGP LGR1 and JGP LGR1, it is a safe extension.

12452 confirmed

MSS to CDNC+(CGP LGR alpha)

197 under review, TBD

China mainland, Taiwan and Hong Kong language experts are reviewing **197** chars and their variant chars, this review might change the variant chars setting of 197 chars, and might affect JGP.

This work could be done in Oct.

12452 confirmed

The next step: 197 chars review

48BC 郱	8262 艪	39F8捛
65FF旿	88B5 袵	39FE捴
732F猯	894D 襟	3A18 揩
9EB9 麹	8B0C 謞	3A52 攃
4C81 鰗	8F19軩	3A67 攪
5227 刧	945A 鑽	3B39 靜
524F翗	9771 靱	3DE7 燺
5605 嘅	984B 顠	3DEB 燬
6060 恈	9DC0 鷀	3E74 狾
6335 拵	34E4 剖	3ED0 玹
656D 敩	3577 咟	4065 矋
681E 栞	35A1 唉	406A 瞮
7460 瑠	35AD 哶	40BB 碾
74A2 瑶	35BF 唭	40DF 礗
74C8 瓈	35CE 喫	4137 秞
750E 甎	35F3嗒	44EA 萏
754A 畊	35FE嘰	4606 蟣
7ADA 竚	3960 惛	47F4踮

The three parties haven't reached consensus on 102 chars, to treat them as independent chars or variant related chars?

The next step: coordination

MSR

86 · 4612 · 36 · 161

7722 · CDNC · 32 · IICore

.CN/.TW/.HK Registered · Normalized

K LGR

197 chars re-coordinate with JGP.

KGP provided K-LGR v0.3 in September, CGP has not compare C-LGR and K-LGR. We will do it after 197 review and re-coordination of C and J.

Thanks

Q&A

Japanese Generation Panel

Hiro Hotta

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)
  - September 29   formal meeting (12)

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 as registry consultant
  - Shuichi Tashiro
    - Character codes
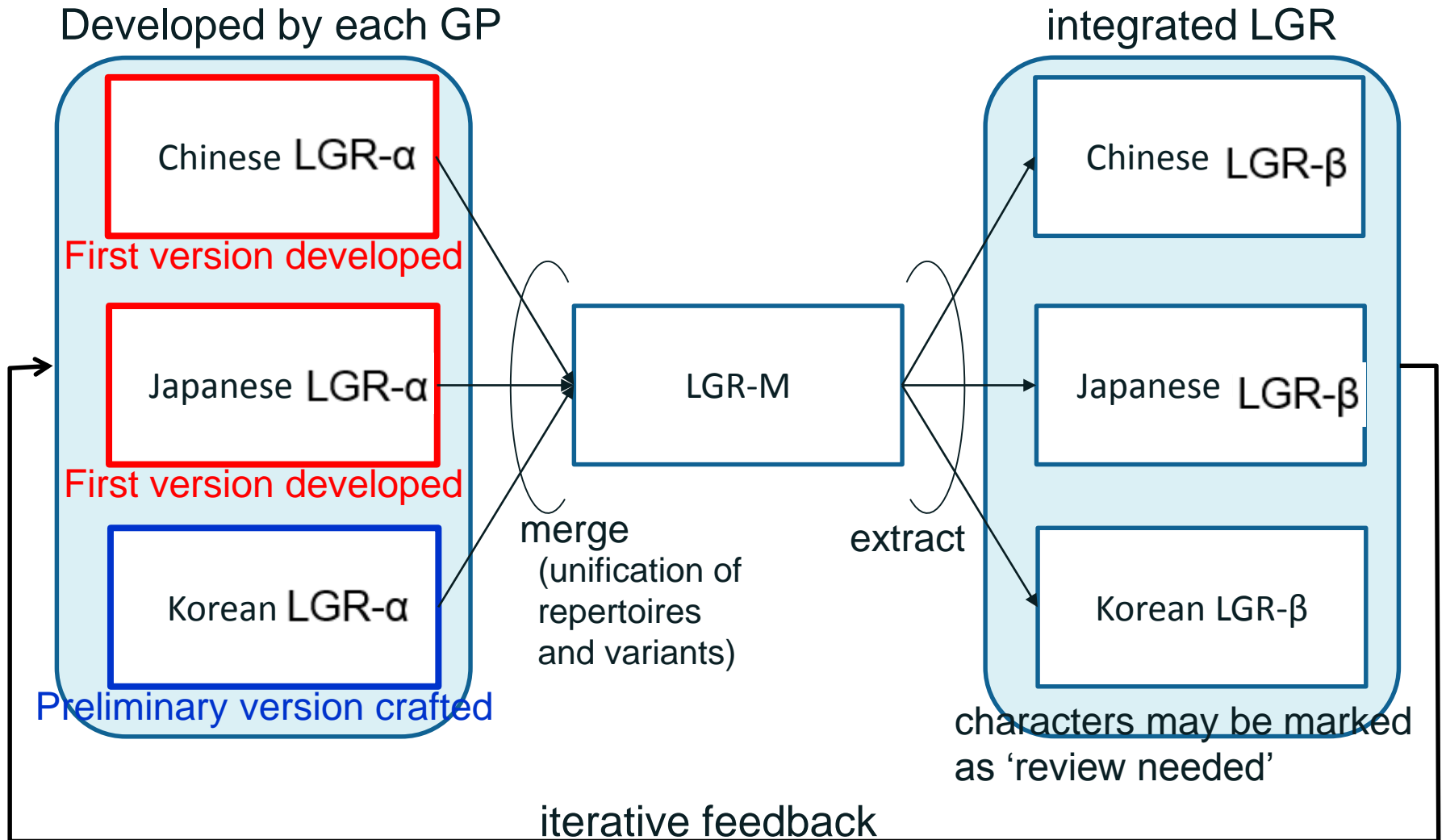 - Yoshiro Yoneya
    - Technical aspects of IDN, LGR

Relationship among CJK language LGRs

Chinese LGR

Japanese LGR

Korean LGR

e.g., 漢

script

e.g., ひ
Hiragana

e.g., ア
Katakana

Han*

e.g., 한
Hangul

...

Japanese GP

Chinese GP

KoreanGP

coordination
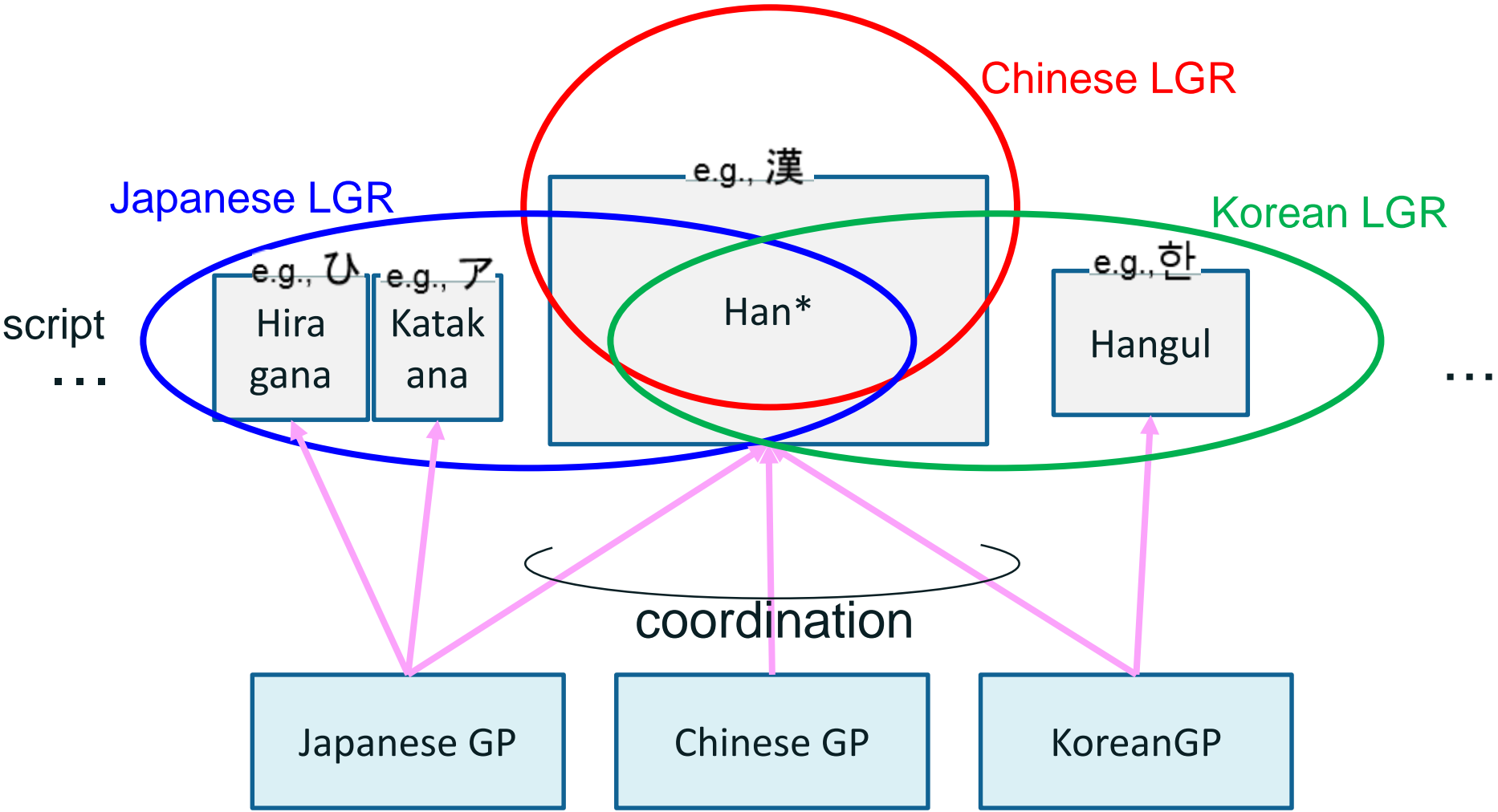
* "Han" is called "Kanji" in Japan, "Hanja" in Korea

Framework of CJK LGR integration for Han characters
(revised by agreement in Buenos Aires)

Developed by each GP

integrated LGR

Chinese LGR-α

First version developed

Japanese LGR-α

First version developed

Korean LGR-α

Preliminary version crafted

LGR-M

merge
(unification of
repertoires
and variants)

extract

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 of LGR development
    - through discussion with ICANN/IP
    - through discussion with CGP and KGP (as well as IP)
    - through investigation inside JGP
- ## Development of Japanese LGR
  - CJK LGR integration algorithm was developed, agreed, and revised by CGP, JGP, and KGP as a framework
  - As the input to the algorithm, preliminary Japanese LGR (which is called LGR-1) was developed

Discussion status for Japanese LGR-1

More detailed explanation is from the next slide

- ## 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-1 will define no variants for itself
    - Integrated Japanese LGR (whick is called LGR-2) will import (= passively adopt) variants of Chinese LGR-1 and Korean LGR-1
    - Types of each variant in Japanese LGR will be defined in a systematic way == > need more investigation if reduction of allocatable labels is needed, and how it can be done if needed

- ## WLE (whole label evaluation)
  - Japanese LGR-1 may have no or very limited number of tiny rules even if defined == > need more investigation

Overview of Japanese LGR-1 (J-LGR-1)

- 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 are not yet finalized (nor for LGR-1 but for LGR-2 after CJK integration)
- WLE
  - Rules (although not very many) are under discussion

Development after Buenos Aires

- Is reduction of allocatable labels really necessary if the reduction harms the Japanese word habitude?
  - Variant labels will exist by importing CGP variant characters, although JGP defines no variants
  - So, study Kanji domain names currently registered under .jp
    - biggest size of the set of variant allocatable labels will be 20,736
    - Biggest size of the set of variant allocatable labels as Japanese domain name will be 540
    - Biggest number of labels that are mutually variants registered under .jp is 4
  - We believe these numbers are small enough to manage
- Labels requested by applicants should be treated allocatable in case very strict rules to reduce allocatable labels is employed
    - In addition, multiple variant labels should be able to be registered when applicant knows which of them are needed for them
  - This demand should be requested to ICANN

Community Driven Way Forward: Task Force on Arabic Script IDNs

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

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

TF-AIDN Objectives:

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

Process Undertaken

- Process undertaken for developing the proposal:
  1. Defining the code point principles.
  2. Finalizing the code point repertoire.
  3. Defining the code point variant principles.
  4. Finalizing the code point variant sets.
  5. Analyzing allocatable vs. blocked variants.
  6. Finalizing code point variant types.
  7. Analyzing Whole Label Evaluation rules.
  8. Finalizing WLE rules.
  9. Documenting the process.
  10. Creating XML LGR for Arabic script LGR proposal.
  11. Finalizing the Documentation and XML LGR proposal for submission to ICANN.
      - https://www.icann.org/news/announcement-2-2015-08-24-en

Code Point Principles

1. Any code point, which is a letter in established contemporary use in a language.
2. Any code point DISALLOWED by IDNA 2008 protocol.
3. Any code point representing a security or stability issue.
4. Any code point not listed in the Arabic GP proposal.
5. Any code point specifically encoded for historic use without established contemporary use.
6. The generation panel lacked sufficient information on the usage.
7. The generation panel could only ascertain the use for such languages that had an EGIDS rating higher than five.
8. The generation panel had data on the use of code points, but where Integration Panel explicitly expressed disagreement on the validity and relevance of such data in separate communications

Repertoire

Analysis of Code Point Variants

1. Visually identical in positions/forms (isolated, initial, medial or final form).
2. Two code points are variants if at least a sub-set of the Arabic script-using community employs those two code-points as stylistic variations of one another.
3. Two code points are variants if they only differ in their dot orientation.
4. Same letter in Western (African) Arabic orthography and Conventional Arabic orthography are considered variants.

Conventional Arabic Orthography	Western (African) Orthography
ف	ڡ
ق	ف
ن	ں
ك	ﯓ
ى	ے

Summary of the Work

Number of code points: 128.

Variants:
Total number of variants: 192 (this is more than the code points as the variants are directional)
Number of variants for type 'allocatable': 26.
Number of variants for type 'blocked': 166.

WLE Rules:
Number of rules defined: 17.

LGR at Work!

For the IDN ccTLD ایران for Iran

Code point sequence: 0627 06CC 0631 0627 0646

Generated Variants: 400

   allocatable=4 (including 1 original)

   blocked=396

Perspectives

Next steps:

- Arabic script LGR for the second level (started)
- Universal acceptance of Arabic IDNs

Armenian language

- The **Armenian language** is an Indo-European language spoken by the Armenians. It is the official language of the Republic of Armenia and the self-proclaimed Nagorno-Karabakh Republic. It has historically been spoken throughout the Armenian Highlands and today is widely spoken in the Armenian Diaspora. The following web site contains a full description of the Armenian repertoire for use in the Armenian writing system:

  http://www.omniglot.com/writing/armenian.htm

Selected code points

Code	Script	Name	Code	Script	Name
0561	ɯ	ARMENIAN SMALL LETTER AYB	0574	մ	ARMENIAN SMALL LETTER MEN
0562	բ	ARMENIAN SMALL LETTER BEN	0575	յ	RMENIAN SMALL LETTER YI
0563	q	ARMENIAN SMALL LETTER GIM	0576	ն	ARMENIAN SMALL LETTER NOW
0564	դ	ARMENIAN SMALL LETTER DA	0577	շ	ARMENIAN SMALL LETTER SHA
0565	ե	ARMENIAN SMALL LETTER ECH	0578	n	ARMENIAN SMALL LETTER VO
0566	q	ARMENIAN SMALL LETTER ZA	0579	չ	ARMENIAN SMALL LETTER CHA
0567	է	ARMENIAN SMALL LETTER EH	057A	պ	ARMENIAN SMALL LETTER PEH
0568	ը	ARMENIAN SMALL LETTER ET	057B	ջ	ARMENIAN SMALL LETTER JHEH
0569	թ	ARMENIAN SMALL LETTER TO	057C	ռ	ARMENIAN SMALL LETTER RA
056A	ժ	ARMENIAN SMALL LETTER ZHE	057D	u	ARMENIAN SMALL LETTER SEH
056B	ի	ARMENIAN SMALL LETTER INI	057E	վ	ARMENIAN SMALL LETTER VEW
056C	լ	ARMENIAN SMALL LETTER LIWN	057F	տ	ARMENIAN SMALL LETTER TIWN
056D	խ	ARMENIAN SMALL LETTER XEH	0580	ր	ARMENIAN SMALL LETTER REH
056E	ծ	ARMENIAN SMALL LETTER CA	0581	g	ARMENIAN SMALL LETTER CO
056F	կ	ARMENIAN SMALL LETTER KEN	0582	ւ	ARMENIAN SMALL LETTER YIWN
0570	h	ARMENIAN SMALL LETTER HO	0583	փ	ARMENIAN SMALL LETTER PIWR
0571	ձ	ARMENIAN SMALL LETTER JA	0584	ք	ARMENIAN SMALL LETTER KEH
0572	ղ	ARMENIAN SMALL LETTER GHAD	0585	o	ARMENIAN SMALL LETTER OH
0573	ճ	ARMENIAN SMALL LETTER CHEH	0586	ֆ	ARMENIAN SMALL LETTER FEH

Variants

The Armenian repertoire does not contain in-script variants.

Cross-script homoglyphs

Cyrillic, Greek and Latin scripts are examples of related scripts where some cross-script homoglyphs exist.

Armenian and Cyrillic homoglyphs

Armenian Letter	Cyrillic Letter
ա U+561 Armenian small letter AYB	ш U+448 Cyrillic small letter SHA
հ U+0570 Armenian small letter HO	һ U+04BB Cyrillic small letter SHHA
օ U+0585 Armenian small letter OH	о U+043E Cyrillic small letter O

Armenian and Greek homoglyphs

Armenian Letter	Greek Letter
η U+0572 Armenian small letter GHAD	η U+03B7 Greek small letter ETA
ι U+0582 Armenian small letter YIWN	ι U+03B9 Greek small letter IOTA
օ U+0585 Armenian small letter OH	o U+03BF Greek small letter OMICRON

Armenian and Latin homoglyphs

Armenian Letter	Latin Letter
q U+0566 Armenian small letter ZA	q U+0071 Latin small letter Q
h U+0570 Armenian small letter HO	h U+0068 Latin small letter H
n U+0578 Armenian small letter VO	n U+006E Latin small letter N
u U+057D Armenian small letter SHE	u U+0075 Latin small letter U
g U+0581 Armenian small letter CO	g U+0067 Latin small letter G
ι U+0582 Armenian small letter YIWN	ι U+0269 Latin small letter IOTA
o U+0585 Armenian small letter OH	o U+006F Latin small letter O

Public comments

Public Comments was opened on Armenia Label Generation Rules proposal

- *https://www.icann.org/public-comments/proposal-armenian-lgr-2015-07-22-en*
- *https://www.icann.org/public-comments/proposal-armenian-lgr-22jul15-hy-2015-07-22-hy*

1. Much of the Armenian Diaspora in the world is concentrated in Russia, the US and France. Russia can be mentioned in the list of countries where the Armenian language is widely spoken…
2. The first letter of the Armenian Alphabet, the letter AYB 'ա' and the Cyrillic letter 'ш' (SHA) can be considered as cross-script homoglyphs. As a symbol and term the Armenian AYB (Ա ա) corresponds to the Greek letter Alpha and is the most frequently used letter of the Armenian language. …

Public comments

2. WIKIPEDIA indeed indentified Russia as one country with a large Armenian Diaspora https://en.wikipedia.org/wiki/Armenian_diaspora (This information is corroborated by figures obtained independently on other Web sites that I checked)

3. In my eyes, Armenian letter AYB 'ա' and the Cyrillic letter 'ш' (SHA) indeed look like cross-script homoglyphs. It is worth noting that pointing out cross-script homoglyphs is important due to the possibility of visual Iconfusion not only by users of the script but indeed every Internet user.

Public comments

4. The Armenian proposal is missing the test cases that the Integration Panel requires for mechanical review of the LGR file. They need to be added to the Armenian proposal to make it complete….

What the Integration Panel is looking for is a set of plain text files containing a few dozen "typical" labels and variant labels that can be evaluated against the repertoire, variants and any whole-label rules. Please include both valid as well as invalid labels, but with an indication of which is which and why a label should be invalid.

According to Public Comments the following changes were made in the Proposal:

-***The list of Armenian Diaspora was updated.***

- ***Armenian letter AYB 'ա' and the Cyrillic letter 'ш' (SHA) were added to the cross-script homoglyphs.***

- ***"Test labels for .հայ" file was provided.***

Current Status

- A sunrise period is announced for .հայ domain on Sept 1, 2015
- A landrush period will follow from Dec 1, 2015
- Regular registrations will start on Jan 1, 2015

Future Plans

Armenian GP is starting to work on **Universal Acceptance** problems (IDN e-mail clients, Internet applications, mobile IDN, etc)

Cyrillic Generation Panel

Yuri Kargapolov and

Dusan Stojicevic

ICANN

Agenda of Cyrillic GP

1 Main Challenges

2 Topics

3 Status of GP

4 Current Achievements

5 The Role of IP

6 Future Tasks

Main Challenges

⊙ The geographical and "jetlag" factor

Main Challenges

⊙ Insufficient number of participants in certain groups, e.g. Central Asia and Mongolia

⊙ Irregularity of studies on different regions
  • The most advanced was the "Balkans" working group
  • Then leveled "Russian-Ukrainian-Belorussian" working group
  • There were problems with the "Central Asia" group, but they are being solved and the participants from Kyrgyzstan, Kazakhstan and Tajikistan was joined

⊙ Two issues still persist:
  • The diversity of languages in the Russian Federation
  • The lack of participants for the "Mongolian" group

Topics

◉ Finish the "**Proposal for the Generation Panel (GP) for the Cyrillic Script Label Generation Ruleset (LGR) for the Root Zone**"

◉ See in our Plan
  - *Development and Finalization Principles Inclusion, Exclusion or Deferral of Code Points* **(2 months)**
  - *Development and Finalization of Rules Variants for inclusion of the Code Points* **(2 months)**
  - *Development and Finalization of Whole Label Rules* **(2 months)**
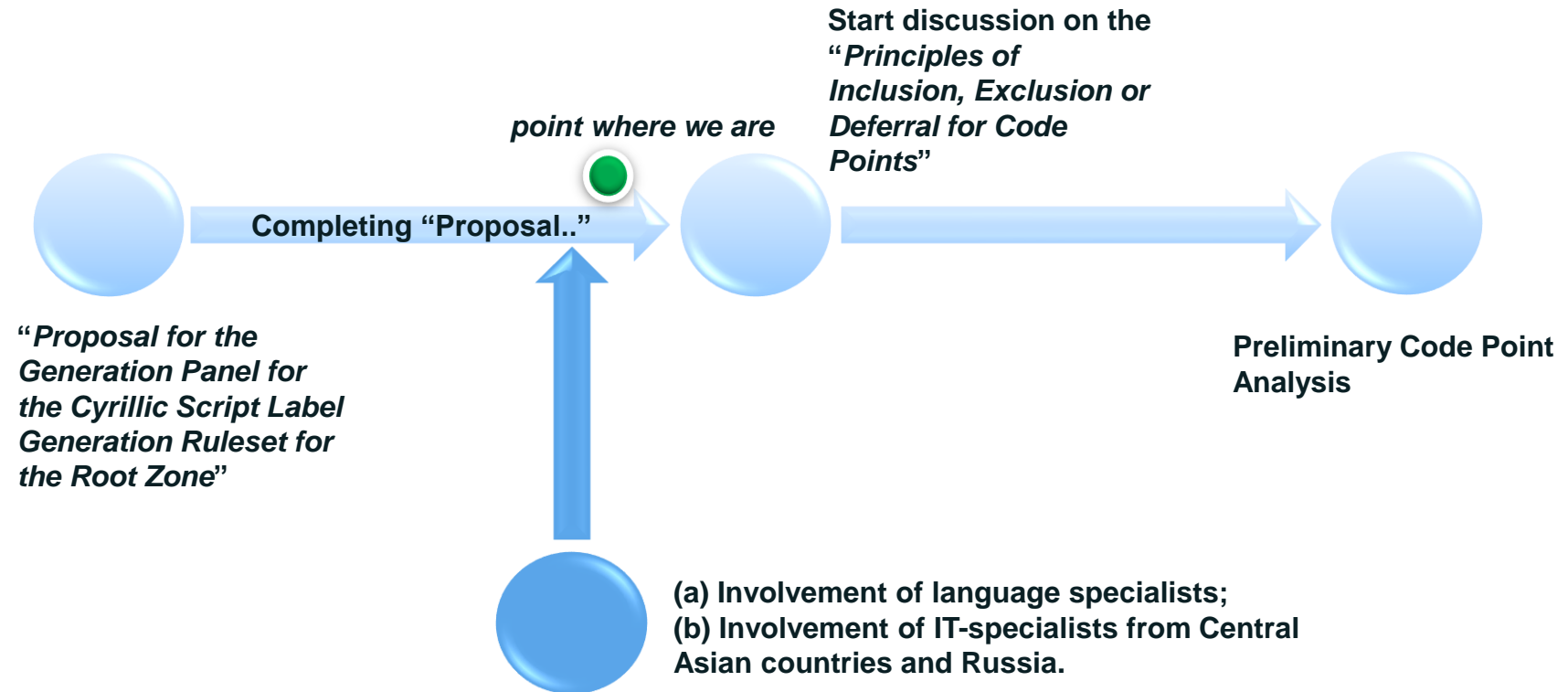 - *Development and Finalization of LGR Documents for Cyrillic Script, and Submission* **(2 months)**

Current Achievements

- Two Cyrillic scripts: Cyrillic - No. 220 Code Cyrl and Cyrillic (Old Church Slavonic variant) – No. 221 Code Cyrs
  Solved. To consideration received one script No. 220 Code Cyrl
- The problem of determining the set of "principal languages using the scripts that should include a number of languages according ISO 639-3"
  Solved. 31 languages on the basis of EGIDS were defined and recorded for further consideration
- We have highlighted a few particular cases that make us look on non-standard in some issues
  Temporarily solved. We temporarily refused to consider these cases at this stage of work. It is necessary to save up more information for cases:
  - Montenegrin mix Cyrillic and Latin letters in one language
  - Ukrainian and Belarus apostrophe letter
  - Uppercase / Lowercase visualization confusion

BIG!
and
Thanks!

Status: Future Tasks

IP: "...Only when the Cyrillic repertoire is fixed (or progressed far enough to settle the status of any candidate homoglyph) can the scheduled work on confusables make headway, and make sense. The Integration Panel therefore suggests that Task 2 (Principles of Inclusion, Exclusion or Deferral for Code Points) precede Subtask 1.2 (Preliminary Code Point Analysis)..."

**Start discussion on the
"*Principles of
Inclusion, Exclusion or
Deferral for Code
Points*"**

point where we are

Completing "Proposal.."

**"*Proposal for the
Generation Panel for
the Cyrillic Script Label
Generation Ruleset for
the Root Zone*"**

**Preliminary Code Point
Analysis**

**(a) Involvement of language specialists;
(b) Involvement of IT-specialists from Central
Asian countries and Russia.**

Thanks!

Questions?

Engage with ICANN and IDN Program

Thank You and Questions

Reach us at: IDNProgram@icann.org
Website: icann.org/idn

twitter.com/icann

gplus.to/icann

facebook.com/icannorg

weibo.com/ICANNorg

linkedin.com/company/icann

flickr.com/photos/icann

youtube.com/user/icannnews

slideshare.net/icannpresentations

ICANN