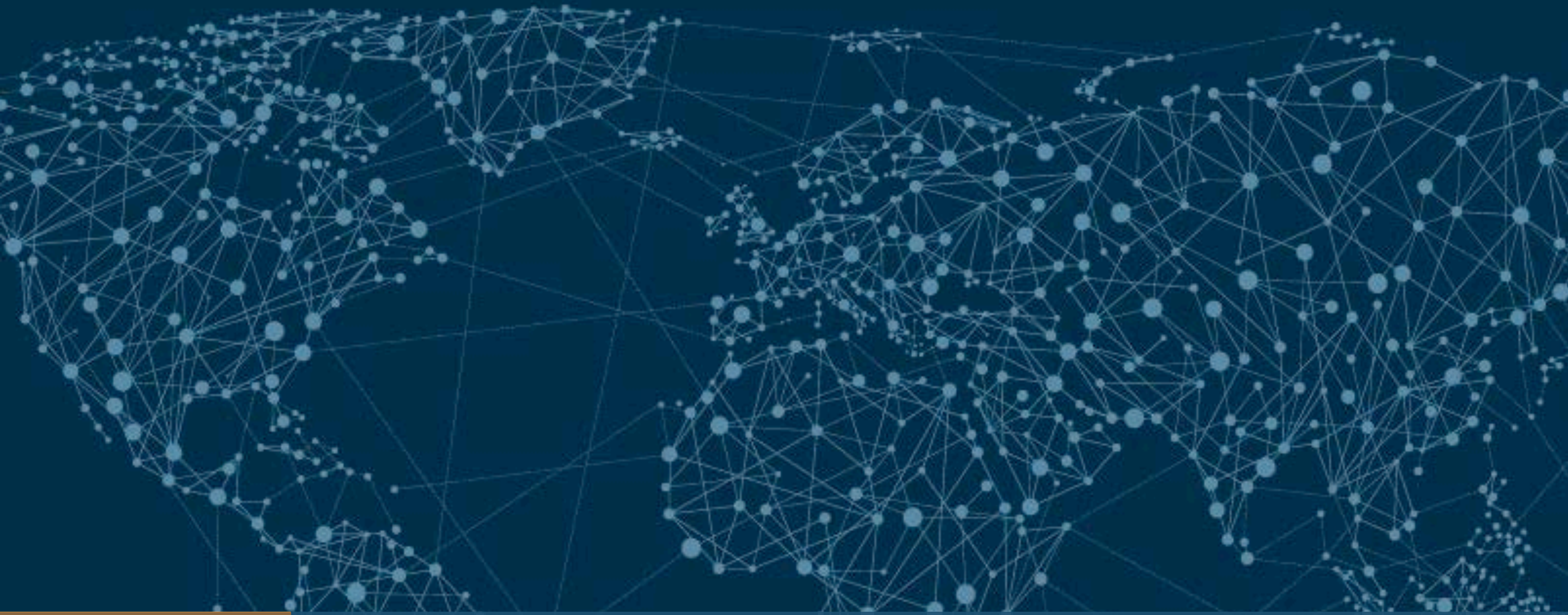
ICANN

# ICANN | 53 Buenos Aires

21-25 JUNE 2015



# Root Zone LGR Workshop

Sarmad Hussain | IDN Program Sr. Manager | 24 June 2015

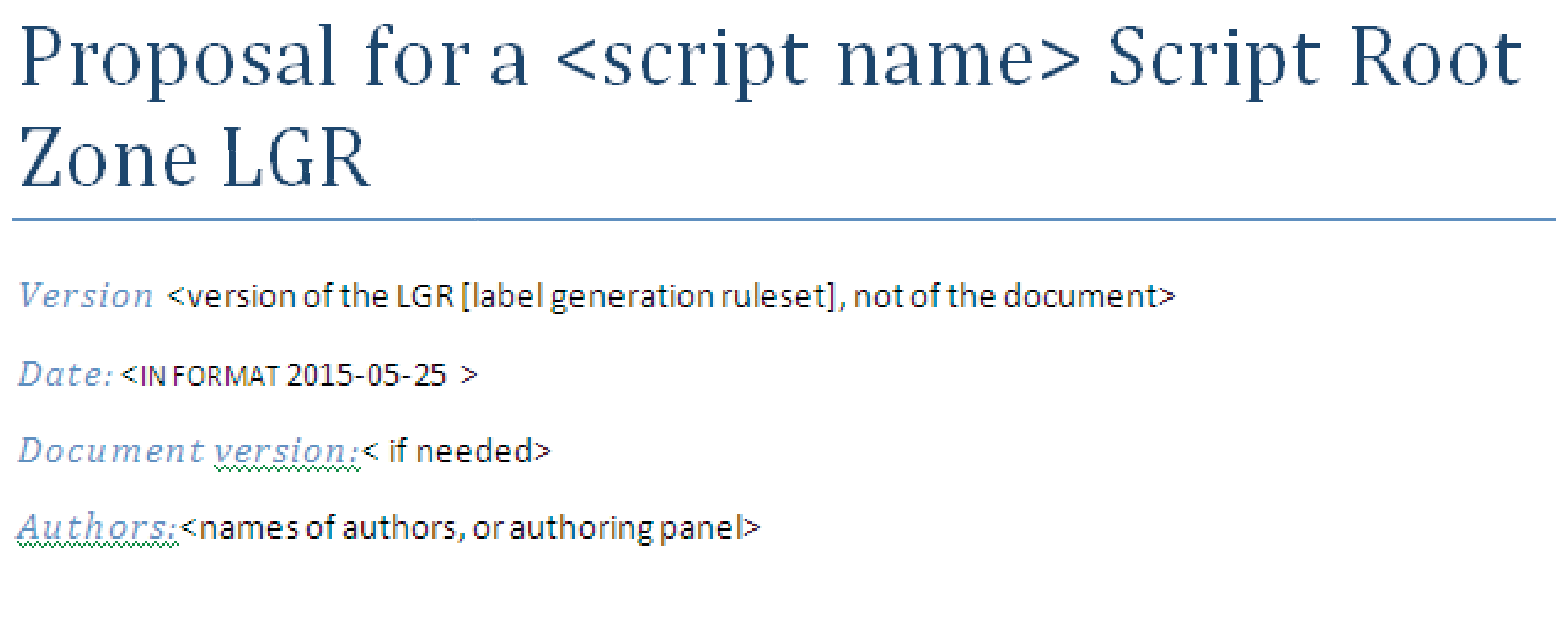
# Agenda

- ⊙ **Requirements of an LGR Proposal** – Nicholas Olster
- ⊙ **LGR Toolset Update** – Marc Blanchet
- ⊙ **Community Updates:**
  - Chinese, Japanese and Korean Coordination – Hiro Hotta
  - Cyrillic Generation Panel – Yuriy Kargapolov
- ⊙ **Q&A**

# Requirements of an LGR Proposal

Nicholas Ostler
Integration Panel
IDN Program
ICANN

# Title and Metadata



# Proposal for a <script name> Script Root Zone LGR

*Version* <version of the LGR [label generation ruleset], not of the document>

*Date:* <IN FORMAT 2015-05-25 >

*Document version:*< if needed>

*Authors:*<names of authors, or authoring panel>

# Required LGR Proposal Elements

1. General Information/Overview/Abstract
2. LGR Proposed Script
3. Background on Script and Principal Languages Using It
4. Overall Development Process and Methodology
5. Repertoire
6. Variants
7. WLE Rules (Whole Label Evaluation)
8. Contributors
9. References

In separate files:

- XML-Format Specification for the LGR
- Test Cases

Remember:

include page numbers and page headers repeating the title.

# LGR Proposals Received So Far

| Template | Armenian | Arabic |
|---|---|---|
| 1. General / Overview | — | [Title and Contents] |
| 2. LGR Proposed Script | 1.1 Script for which the panel is to be established | 1. Script and languages covered |
| 3. Background on Script and Principal Languages Using It | 1.2 Principal languages using that script<br>1.3 Geographic territories or countries with user communities<br>1.4 Related scripts? | APPENDIX A. Some of the languages using Arabic Script |
| 4. Overall Dev. Process & Methodology | 3. Work plan | 2. Process undertaken for developing the proposal – team, repertoire, variants |
| 5. Repertoire | 4.1 Repertoire,<br>APPENDIX 1. MSR-2 | 3. Code point repertoire – summary, included, excluded |
| 6. Variants | 4.2 Variants | 4. Variants – initial analysis<br>5. Variants – final recommend |
| 7. Whole Label Evaluation (WLE) Rules | [4.3. Cross-script homoglyphs] | 6. Whole Label Evaluation (WLE) Rules |
| 8. Contributors | 2. Proposed initial composition of panel | APPENDIX B. Members of the Task Force on Arabic Script |
| 9. References | … | … |

# 1. General Information/Overview/Abstract

This section of an LGR proposal is intended to summarize some of the salient facts about the LGR.

It also gives the links to the auxiliary files.

Example:

> *This document presents the LGR proposal for XXXX script, documents the design methodology and gives the justification for each of the design decisions.*
>
> *It is accompanied by <Proposed-LGR-XXXX.xml>, giving the LGR specification in XML format and Valid-Labels-XXXX.txt and Invalid-Labels.txt containing test cases demonstrating the features defined in the LGR.*

# 2. LGR Proposed Script

Each LGR is for a single script (in the sense of ISO 15924).

Each LGR proposal needs to unambiguously identify the script to which it applies. In the XML, this is done with an RFC 5646 language tag, with the language field set to “und-”. In the background document, copying the ISO 15924 information would be useful, as has been done here for Armenian.

***ISO 15924 Code:  Armn***  
*ISO 15924 Key N°: 230*  
*ISO 15924 English Name: Armenian*  
*ISO 15924 Date: 2004-05-01*  
*English transcription of native script name: Hye*  
*Native name of the script: Հայ*

# 3. Background on Script & Main Languages

Should document use of the script in domain name context.

For example:

- Very schematic background on languages supported by the script, their number, diversity, historical links and geographic distribution
- Just enough to set the context and to identify natural classes of languages, as reflected in writing system
- What issues come into play?
  - In-script variants?
  - Cross-script variants?
  - Others?
- How difficult is it to get reliable information about language use?

# 3. Background on Script & Main Languages

This is the place to make GP's policy clear on:

- **Which languages using the script are to be covered?**
  *Ultimately*: Will any users of these languages need Internet labels?

  - Use of Ethnologue EGIDS scale in relevant languages
    - 4 [Educational] or better
  - Other indicators of significant use (in languages, communities)
    - Presence on Internet, mass media, govt. reforms...
      e.g., Specifically in Arabic:
      - Indiv. languages may affect preference among variants
      - As far as poss., a unitary standard is best for root

- **Nature of evidence: Independent, should go beyond Unicode, MSR**
  - At least, http://www.omniglot.com/writing/
  - Ideally, > Standards documents (for content)
    > Internet presence (for status)

# 4. Development Process and Methodology

1. Overall:

- GPs progressively develop a series of fuller LGR Proposal drafts

- It is imperative that GPs keep a clear log of changes:
  - In successive versions of the proposal and submitting to the IP

- IP can help with content in informal discussion phase, but GP will own the content submitted

- IP cannot assure GPs of acceptance in advance of submission

# 4. Development Process and Methodology

2. Will include development of XML document:

- Purpose
  - To specify the content of the proposal (formal, normative)
- Format
  - Defined in [XML-LGR], as explained in documents [Variant Rules] and [WLE-Rules]
- Help available from background documents
  - Tutorial presentations from previous conferences (see links)
  - Examples – [https://github.com/kjd/lgr](https://github.com/kjd/lgr) (Greek, Thaana)
- Help also available from IP during informal discussion phase
- Need for independent check from Test Cases
- Ultimate ownership
  - GP will produce the final version, to its own satisfaction

# 5. Repertoire

## A listing in readable form, with relevant discussion.

For example:

- For alphabetic, abjad or short syllabic repertoires:
  - The chart pages from the MSR charts are annotated to show adjustments made for the repertoire;

- For long repertoires:
  (Viz logographic scripts (e.g. *Han*), or syllabaries where every combination of elements is separately coded (e.g. *Ethiopic*, or *Korean hangŭl*)
  - A summary, with the full tables added in as an appendix

Important to cite authority for each code point authorized, including (if necessary) language and its vitality score.

# 5. Repertoire

Important to cite authority for each code point authorized, including (if necessary) language and its vitality score. Arabic example (as prepared by TF-AIDN):

| item # | Unicode | Glyph | Code Point name and properties | Languages | EGIDS | Reference / Comments |
|---|---|---|---|---|---|---|
| 1 | 0621 | ء | 0621;ARABIC LETTER HAMZA;Lo;0;AL;;;;;N;ARABIC LETTER HAMZAH;;;; | Arabic, Urdu, Punjabi, Sindhi | 1 Arabic | RFC 5564 |
| 2 | 0622 | آ | 0622;ARABIC LETTER ALEF WITH MADDA ABOVE;Lo;0;AL;0627 0653;;;;N;ARABIC LETTER MADDAH ON ALEF;;;; | Urdu, Malay, Punjabi, Kashmiri, Sindhi | 1 Urdu | RFC 5564 |
| 3 | 0623 | أ | 0623;ARABIC LETTER ALEF WITH HAMZA ABOVE;Lo;0;AL;0627 0654;;;;N;ARABIC LETTER HAMZAH ON ALEF;;;; | Arabic, Malay, Torwali | 1 Arabic | RFC 5564 |
| 81 | 06AE | ڮ | 06AE;ARABIC LETTER KAF WITH THREE DOTS BELOW;Lo; 0;AL;;;;;N;ARABIC LETTER CAF WITH THREE DOTS BELOW;;;; | L'Alphabet National du Tchad (ANT) | 1 ANT | ANT (Alphabet National du Tchad) nat. standard for Chad; Figs in JTC1/SC2/ WG2 N3882 (pp. 19-20) |

# 6. Variants

Will be stated here in human-readable form – e.g., as annotated tables (XML file will carry equivalent coded in XML)

In XML:
Reflexive, Symmetric and Transitive set of mappings among code points.

- This requirement is to ensure full coverage in XML file;
- The assignment of Variant types within a mapping is free, and not required to be symmetric or transitive (notably: allocatable, blocked)
- Must co-ordinate other scripts/languages if repertoires overlap (hitherto, just Chinese-Japanese-Korean)

# 5. Variants - Complications

Traditions of different languages may cause need for variants within a script.

For example: **In Arabic script** (a single script for Unicode) languages of different regions may choose different variants:

| Language | feh | qaf | kaf | heh | yeh |
|---|---|---|---|---|---|
| Arabic | ف | ق | ك | ه | ي |
| Persian | ف | ق | ک | ه | ی |
| Urdu | ف | ق | ک | ھ ہ | ی ے |
| Hausa | ڢ | ڧ | ك | ه | ي |

# 6. Variants

Minimize allocatable variant CPs (or else too many alternative labels!)

How to reduce labels generated by variants?

1. **LGR-specific types** (governed by WLE rule) may constrain co-occurrence of variant characters

Chinese WLE example:
“Simp. & Trad. types never co-occur in a label”
(poss.) Arabic WLE:
“African types never co-occur with non-African”

2. Arbitrarily-specified (at registration)

Japanese example:
慶応大学 but also 慶應大学 “Keio Univ.” but
國學院大學 [*not* 國學院大学] “Kokugakuin Univ.”

# 6. Variants – Across Scripts?

Homoglyphs: What script does a (single) character-form belong to?

In scripts of Western origin (e.g., Latin-Greek-Cyrillic-Armenian), some common characters share forms: → cross-script confusables

For example:

- As between Latin & Cyrillic: a e o p c x
- As between Greek & Armenian: *η ι ο*
- As between Latin & Armenian: h n u o

A label composed wholly of such characters (e.g., .ooo, .pea, .hun) can be registered in any script where these forms are possible.

The Armenian Proposal contains a list of such shared glyphs (§4.3)

# 7. Whole Label Evaluation Rules (WLE)

Any other general constraints on the incidence of code points in TLD labels will be stated here in human-readable form (duplicated in XML file for coded equivalent).

Some important notes:

- All rules apply universally in the root (i.e., to all scripts): therefore beware wider implications of those designed for single script
- LGR-specific features (e.g. Chinese Simplified, Traditional, Both), although assigned in Variants section, will require further WLE rules to limit co-occurrences of code points

# Other Items to Conclude the LGR Proposal

8

### Contributors

Brief identifications, with relevant experience, of the scholars who contributed the the Proposal.

9

### References

Full references to authorities cited, including sources of code point information.

10

### Appendices

These items might include longer data tables and other miscellaneous background information.

Further technical files, such as the XML file, and a log file of test results for it, should be submitted separately from the LGR Proposal document itself.

This concludes our guidance on the minimum contents for a successful LGR Proposal.

You don't need more.

# From the Integration Panel

For more details, contact us:

Wil Tan, Marc Blanchet, Asmus Freytag, Michel Suignard, Nicholas Ostler

# Development Process - Resources

- Guidelines for Developing Script-Specific Label Generation Rules for Integration into the Root Zone LGR
  https://community.icann.org/download/attachments/43989034/Guidelines-for-LGR-2014-12-02.pdf

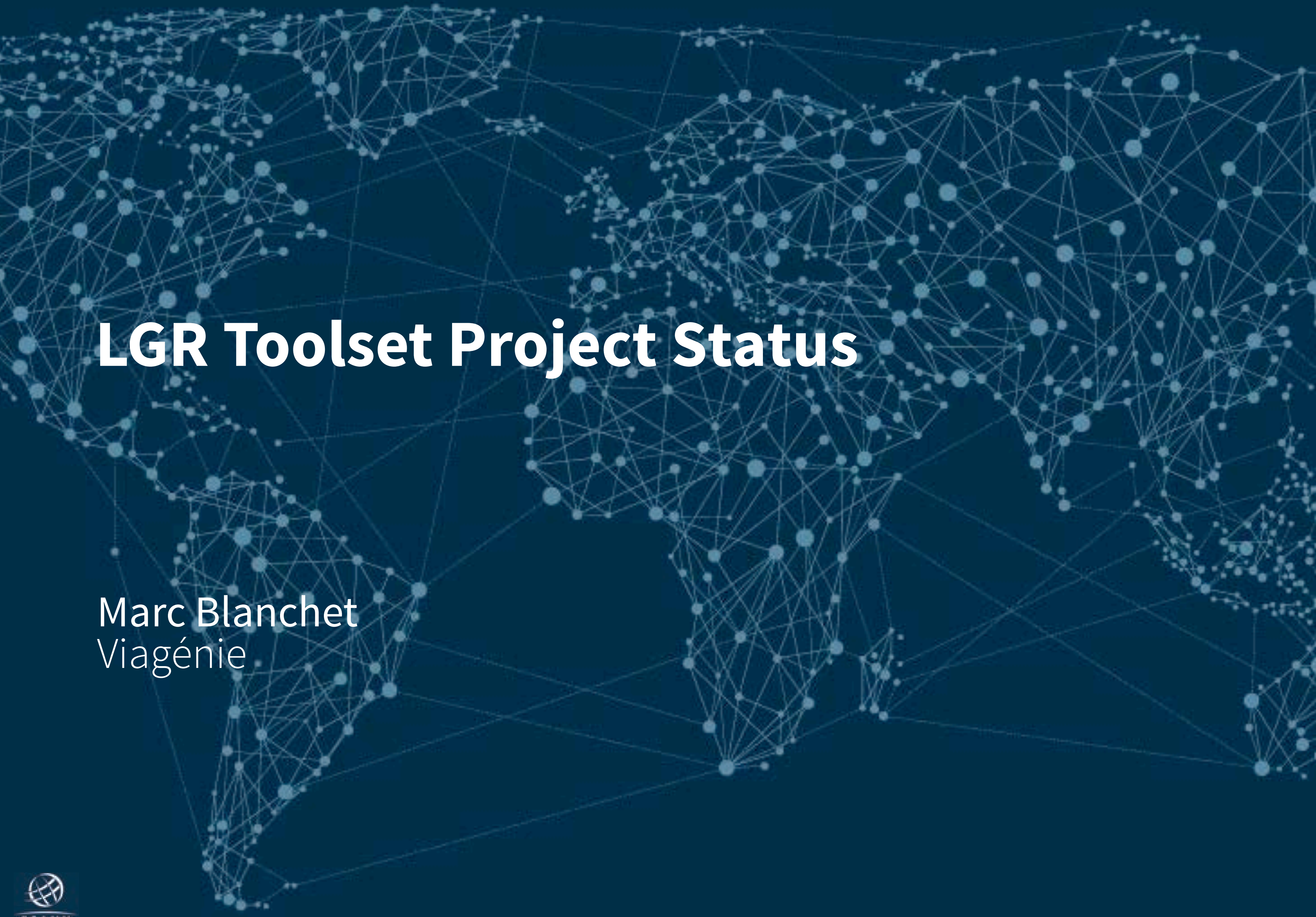
- Variants rules
  https://community.icann.org/download/attachments/43989034/Variant%20Rules.pdf

- Whole Label Evaluation (WLE) Rules
  https://community.icann.org/download/attachments/43989034/WLE-Rules.pdf

- Requirements for LGR Proposals
  https://community.icann.org/download/attachments/43989034/Requirements%20for%20LGR%20Proposals.pdf

- Thaana LGR example
  https://github.com/kjd/lgr/blob/master/resources/Sample-LGR-Thaana.xml

- Greek LGR example
  https://github.com/kjd/lgr/blob/master/resources/Sample-LGR-Greek.xml



# LGR Toolset Project Status

Marc Blanchet
Viagénie

1

Background

2

Project Plan and Timeline

3

UX Design

4

High-level Architecture

5

Current Status

6

Conclusion

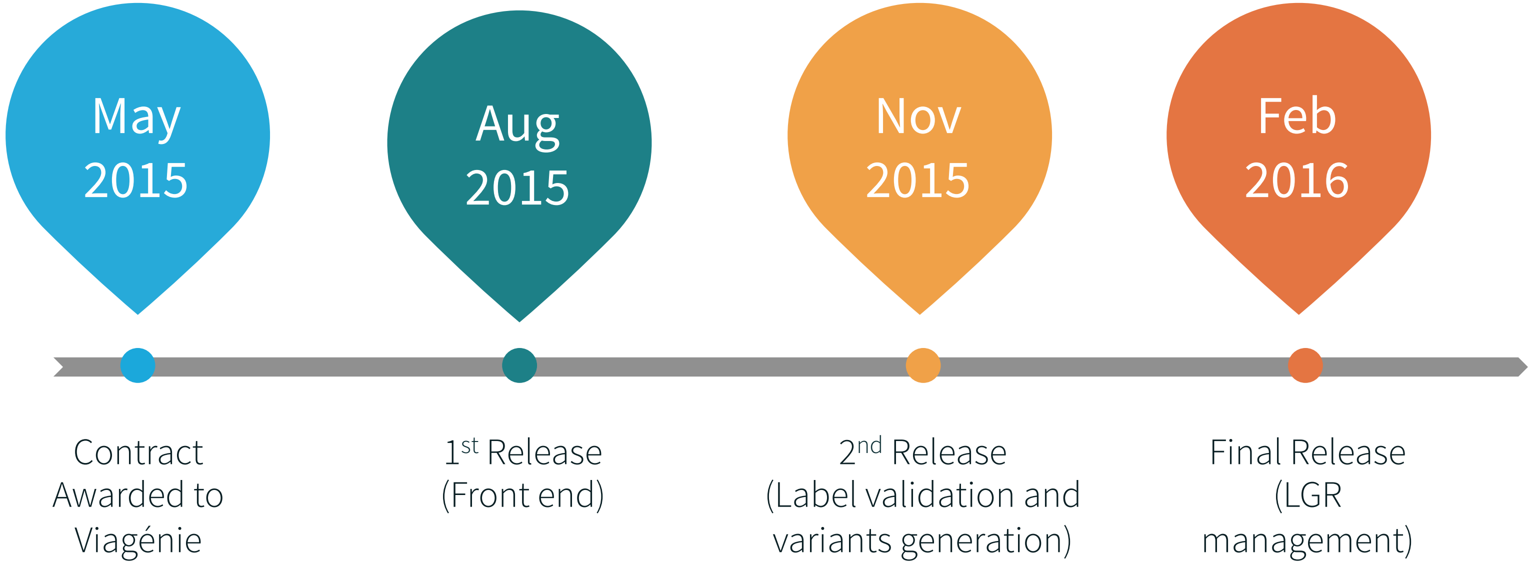
# Background

- LGR XML format can be complicated for some use cases and is cumbersome for non-XML savvy people
- File format does not provide by itself:
  - XML syntax verification
  - LGR XML language verification
  - validation of code points, rules
  - variants specifics (transitivity, symmetry, …)
  - testing of labels
  - etc.
- Project is to:
  - Develop a toolset for LGR
  - Web front-end and CLI
  - Libraries
  - Open-source

# Project Plan

- Three Phases:
  1. LGR creation: web-based application
     - which allows users to construct a LGR through an interactive interface,
     - and have the ruleset export into a syntactically-valid LGR in the XML Format.
     - define the general metadata and linguistic content, including the eligible code points, code point variants and variant dispositions.
  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

# Timeline



May 2015

Contract Awarded to Viagénie

Aug 2015

1st Release (Front end)

Nov 2015

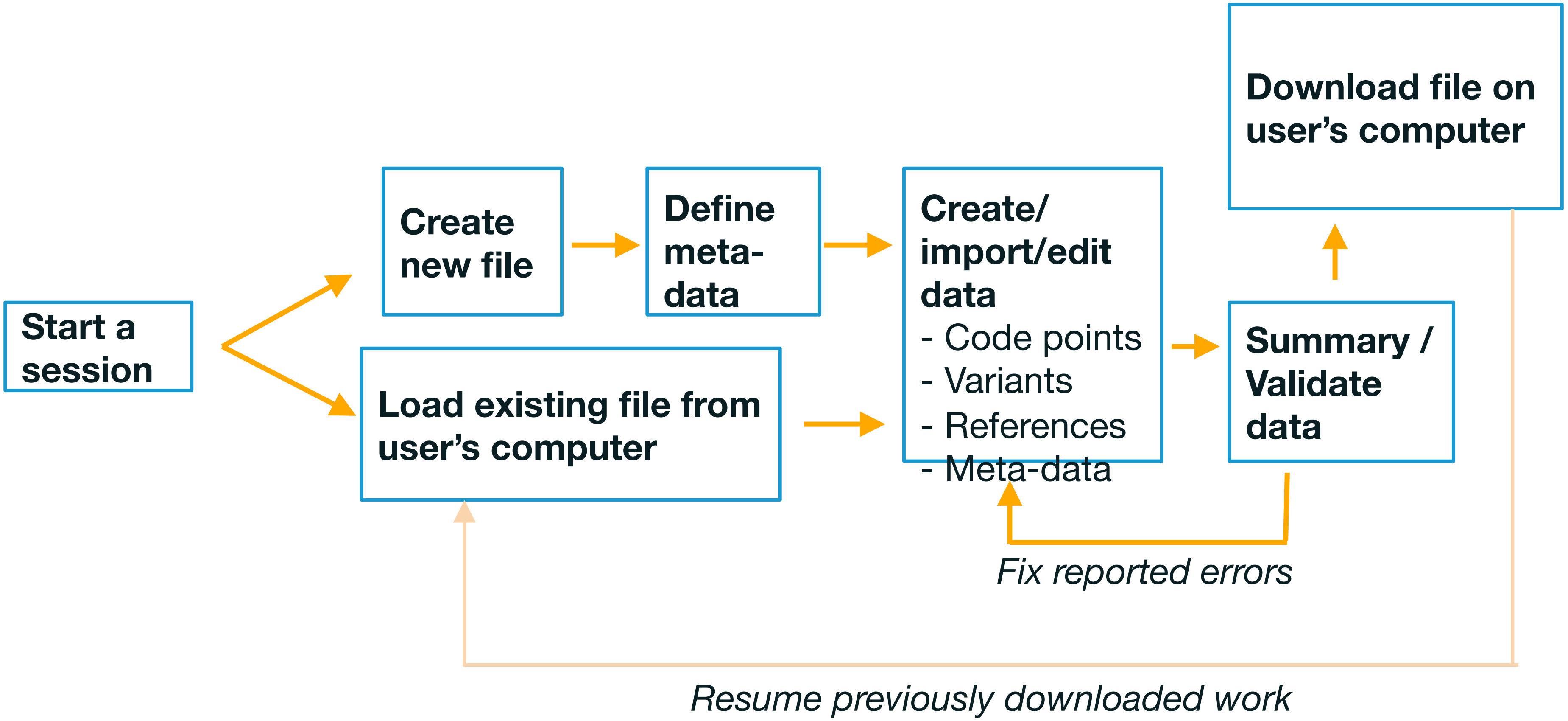
2nd Release (Label validation and variants generation)

Feb 2016

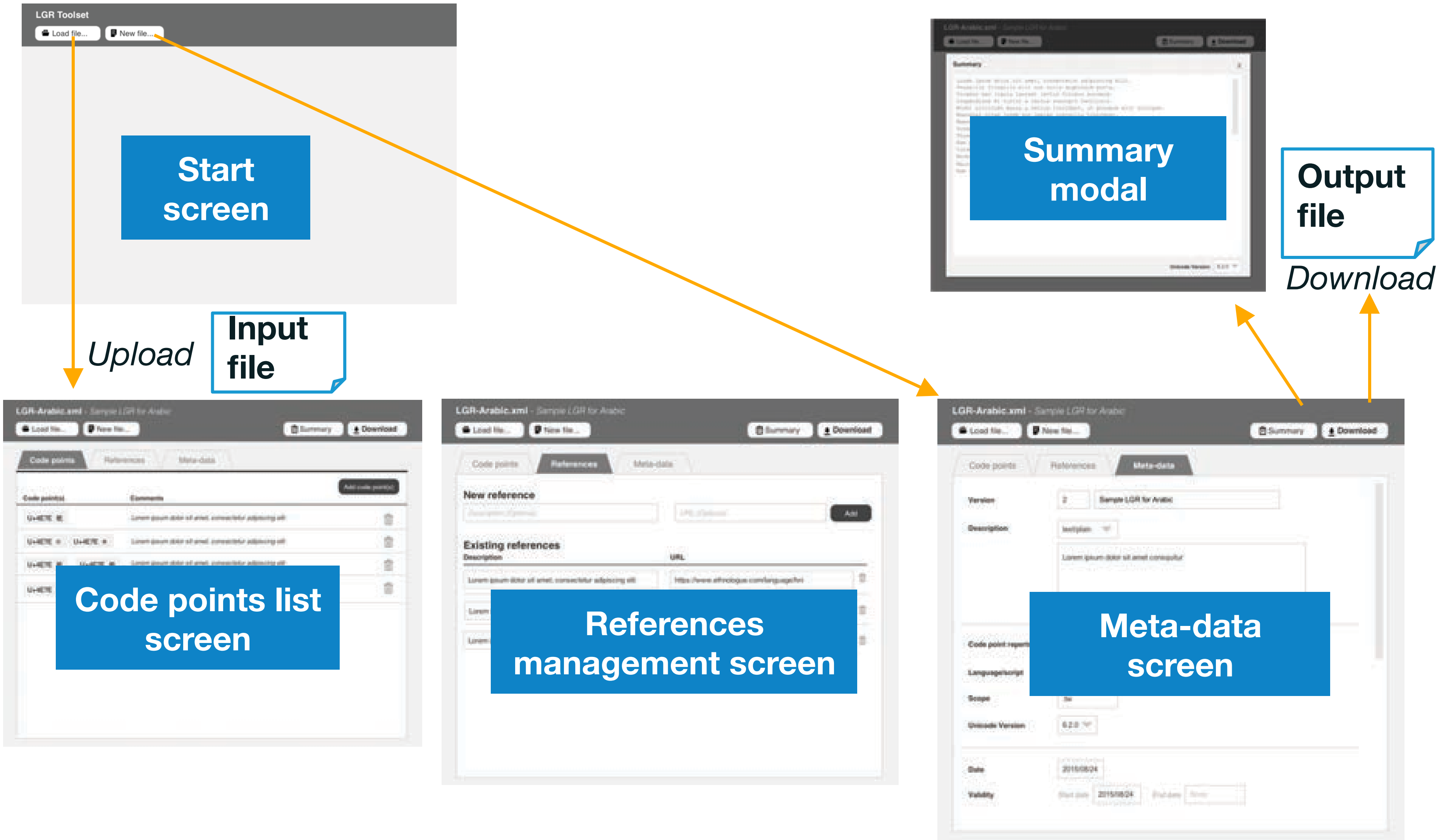
Final Release (LGR management)

# UX Design

# User Work Flow



# Primary Screens Flow



Start screen

Upload

Input file

Code points list screen

References management screen

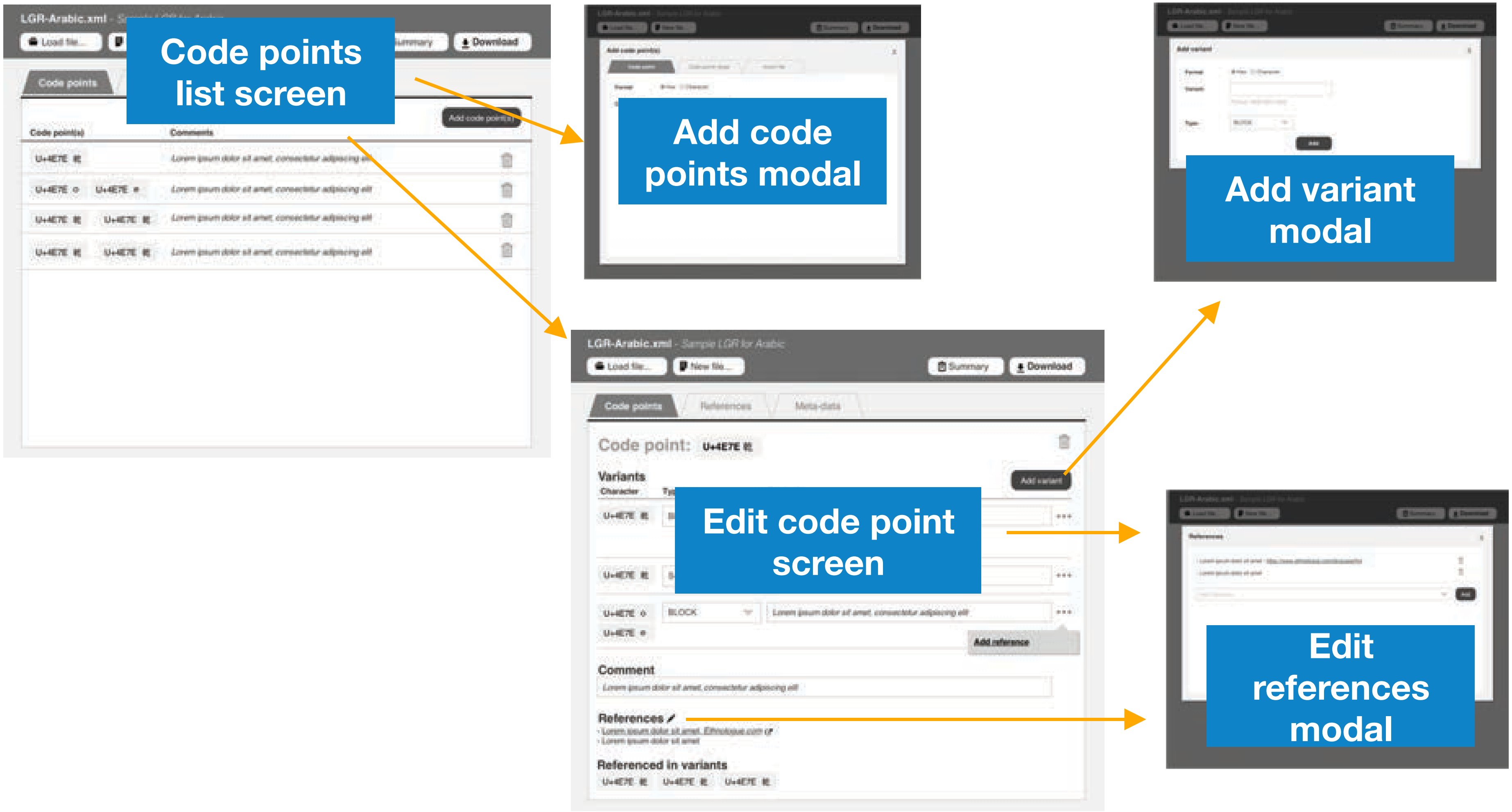
Meta-data screen

Summary modal

Output file

Download

# Code Point Screens Flow



Code points list screen

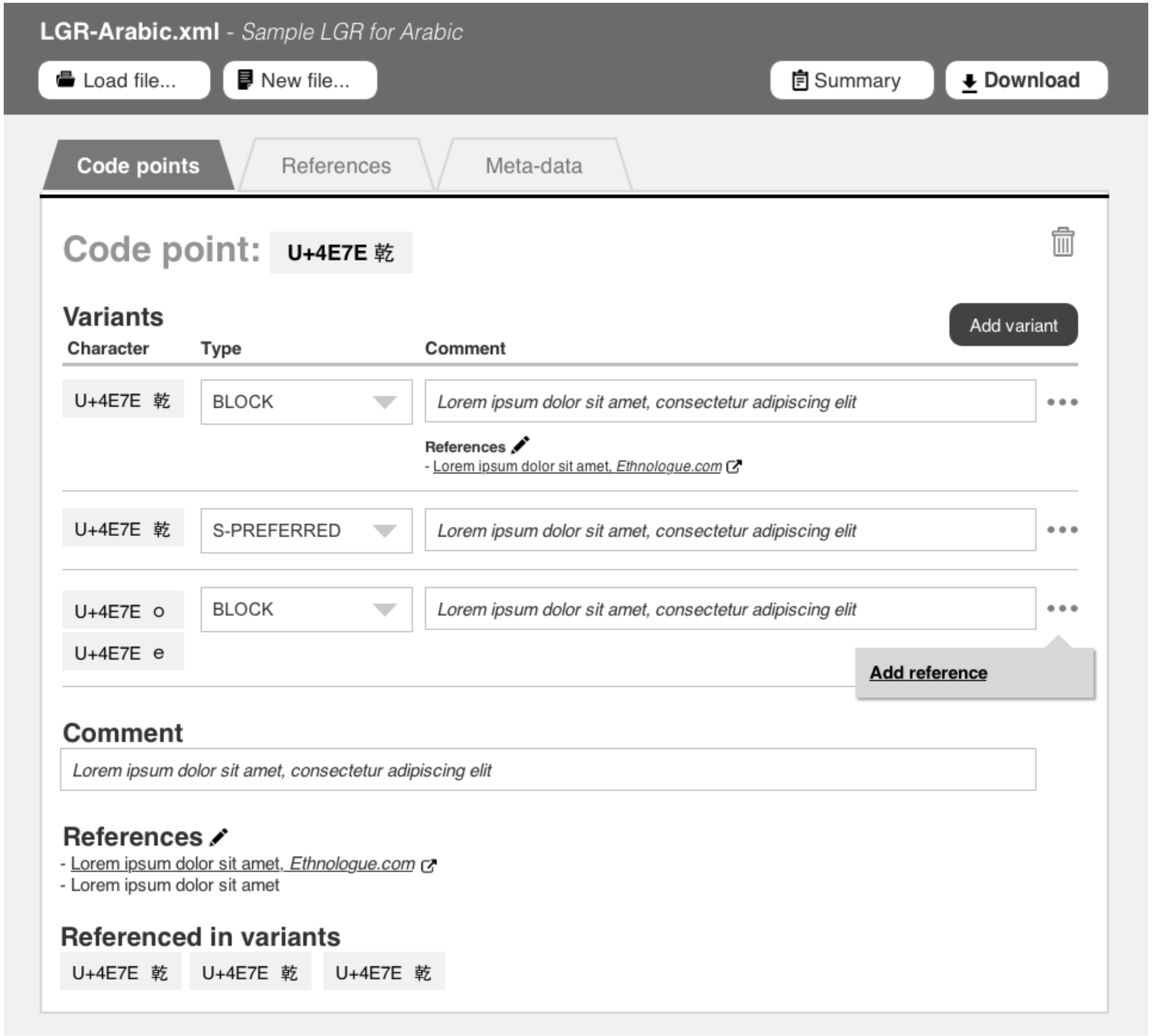
Add code points modal

Add variant modal

Edit code point screen

Edit references modal

# Code Point Editing



**LGR-Arabic.xml** - *Sample LGR for Arabic*

Load file... | New file... | Summary | **Download**

**Code points** | References | Meta-data

## Code point: U+4E7E 乾

### Variants

Add variant

| Character | Type | Comment |
|---|---|---|
| U+4E7E 乾 | BLOCK | *Lorem ipsum dolor sit amet, consectetur adipiscing elit*<br>**References**<br>- Lorem ipsum dolor sit amet, *Ethnologue.com* |
| U+4E7E 乾 | S-PREFERRED | *Lorem ipsum dolor sit amet, consectetur adipiscing elit* |
| U+4E7E o<br>U+4E7E e | BLOCK | *Lorem ipsum dolor sit amet, consectetur adipiscing elit* |

**Add reference**

### Comment

*Lorem ipsum dolor sit amet, consectetur adipiscing elit*

### References

- Lorem ipsum dolor sit amet, *Ethnologue.com*
- Lorem ipsum dolor sit amet

### Referenced in variants

U+4E7E 乾 | U+4E7E 乾 | U+4E7E 乾

# High-level Architecture

# Platform

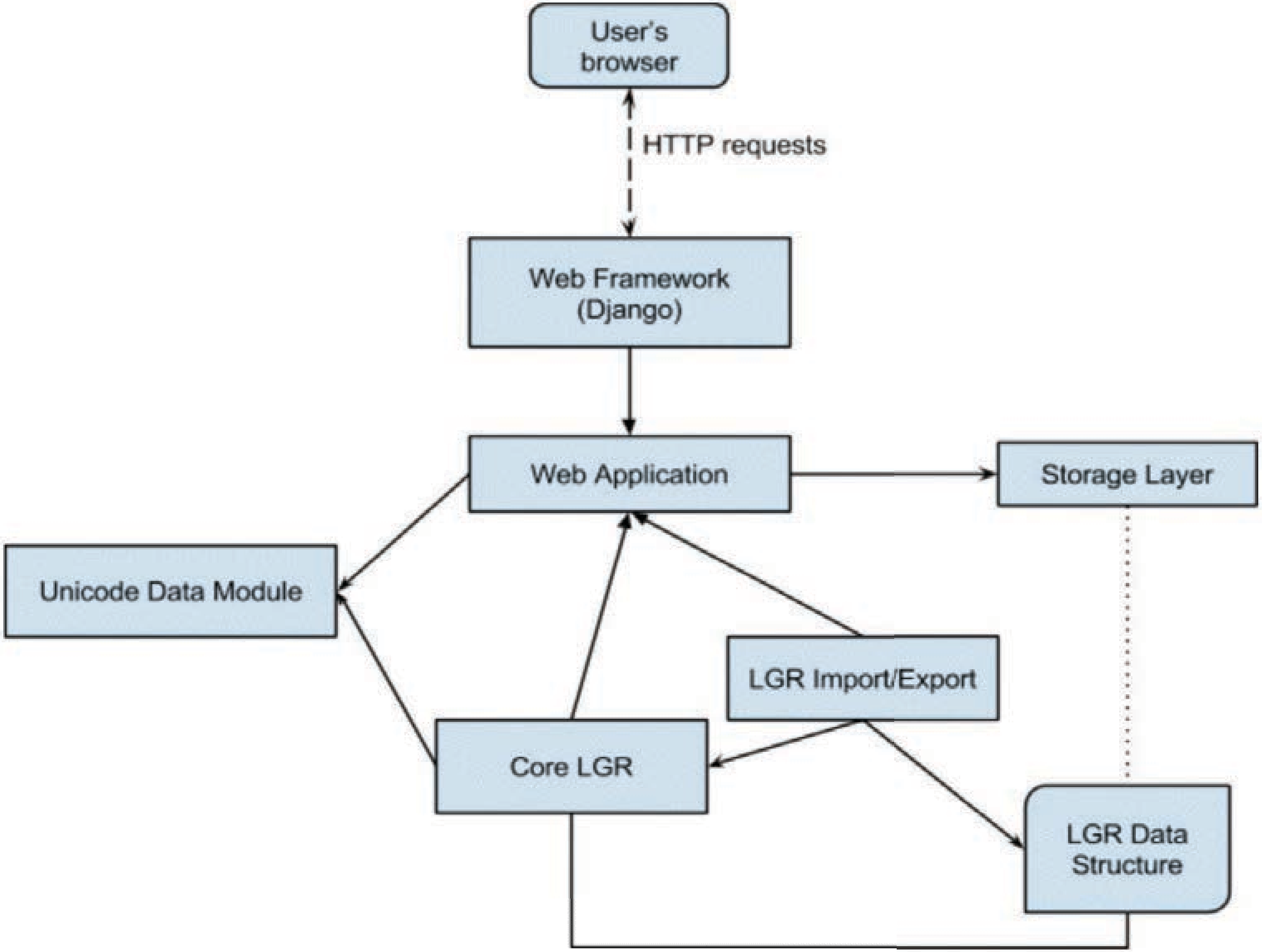
- Deployment (server-side) platform:
  - Linux 64bits
  - Python 2.7.x
  - Django latest stable
  - ICU library latest stable

- Client platform
  - Any modern browser
  - No additional plugin or else needed

# Key Considerations

- No authentication but session management
  - No need for creating/managing a username/password/email
  - Server manages the session
- File editor concept:
  - Only session state is kept on the server
  - Users import and export LGR XML file in each session
- A lot of underlying librairies and code
  - Can be reused for other means

# High-Level Architecture



# Current Status

- Core libraries being written
- First screens being implemented
- Able to import and export
- View LGR
- Some basic validations

- On target for august delivery

- Early august, we will need beta testers. If interested, please contact me directly (mailto: marc.blanchet@viagenie.ca)

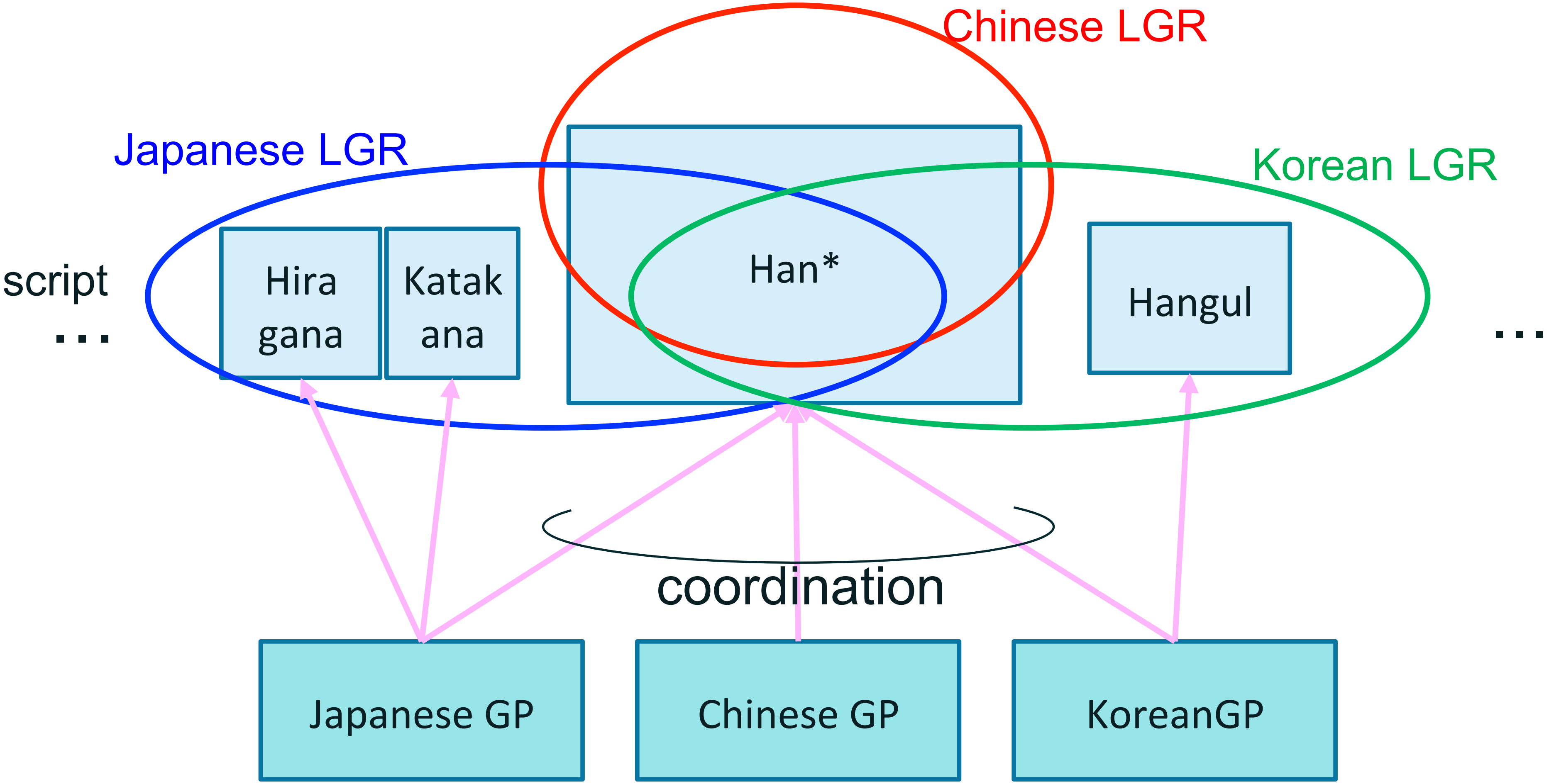
# Conclusion

- LGR work is complex and not user-friendly for non-technical people
- Toolset will help streamline the process to create and manage LGRs
- Open-source will help other entities to use and contribute and enhance the toolset.
- Looking for beta-testers for early august testing (mailto: marc.blanchet@viagenie.ca)

# CJK Coordination Challenges and Solutions

Hiro Hotta JGP chair
Wang Wei CGP co-chair
Kenny Huang CGP co-chair
Kim Kyongsok KGP chair

# Relationship among CJK language LGRs



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

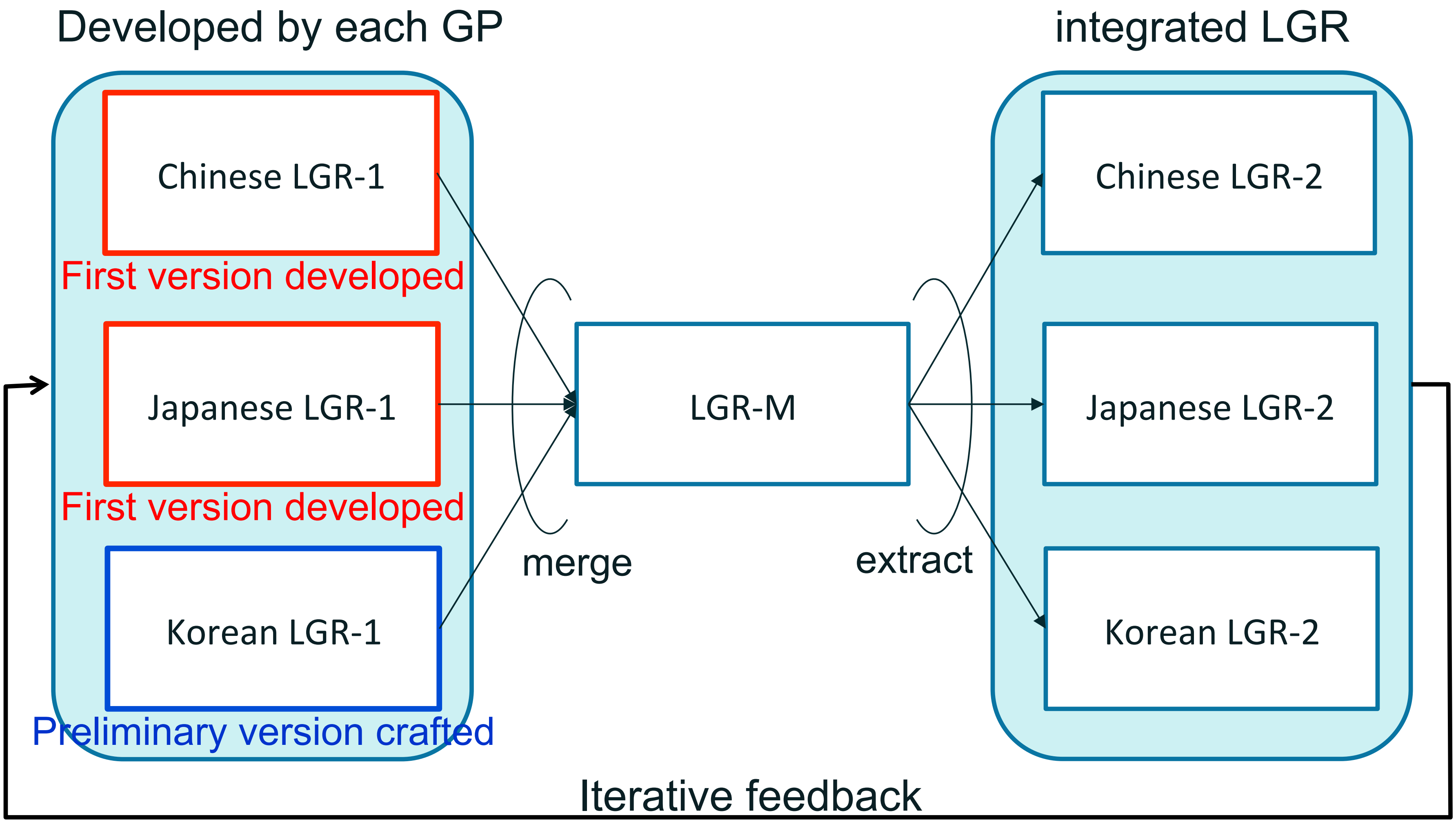
# Typical Issues (especially re. Han characters)

- Each of CJK has thousands of Han characters
  - MSR has about 20,000 Han characters
  - CGP picks up about 13,000 Han characters from MSR
  - JGP picks up about 6,000 Han characters from MSR
  - KGP picks up about 6,000 Han characters from MSR
- Many Han characters are shared by CJK
- Some characters have different usage/meaning in different languages
- Variant definition is different in different languages
  - CGP defines about 3,000 variant groups (e.g. 国&國、机&機)
  - JGP defines no variants (all characters are independent)
  - KGP tentatively defines 66 variant groups
  - Rules for strings are different from language to language
  - Some combination of characters are prohibited in Chinese strings
  - All combination of characters are allowed in Japanese strings

# CJK Coordination

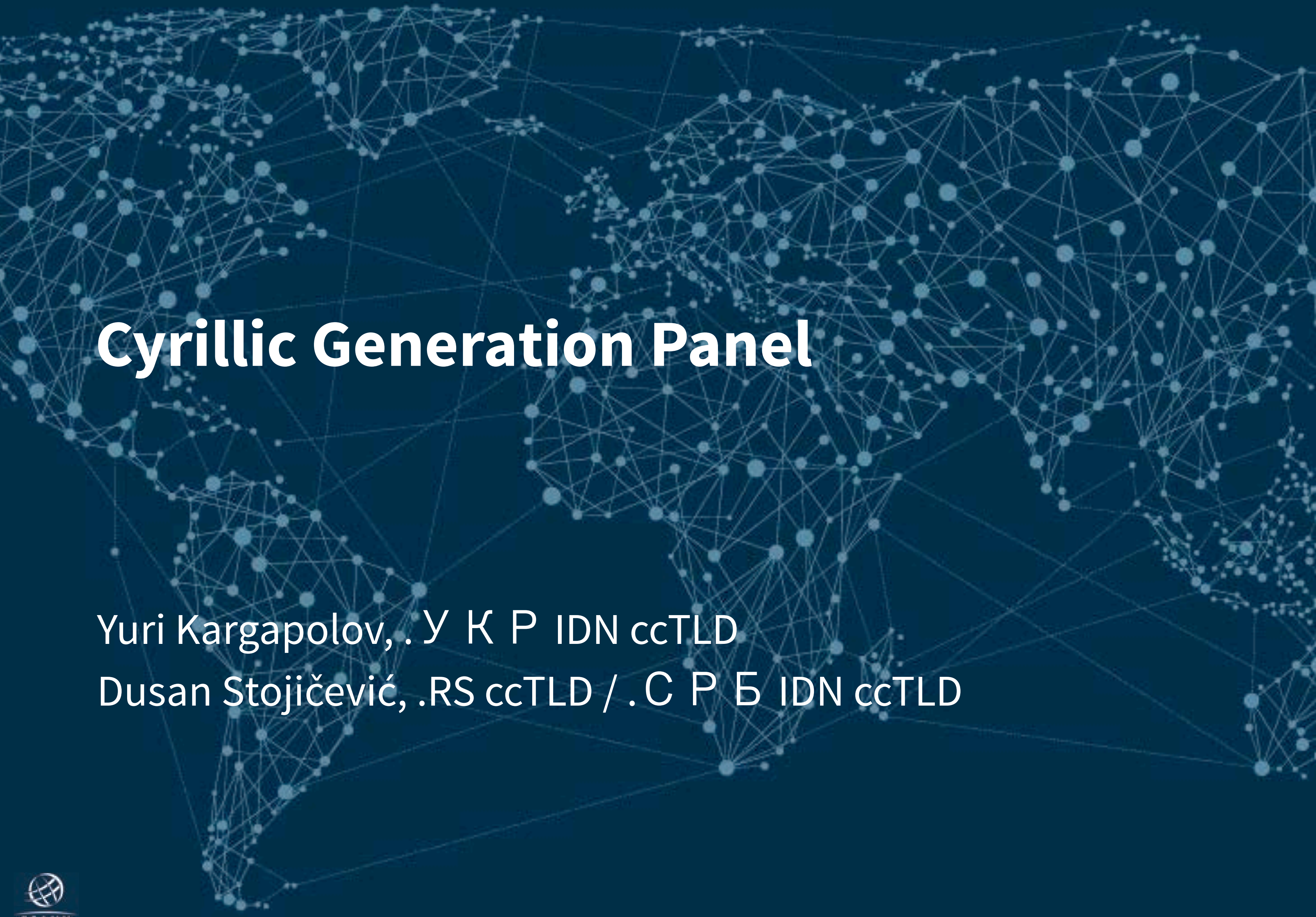
- Ad hoc meetings
  - CGP, JGP and KGP met in ICANN meetings in 2014 and 2015
  - CGP and JGP met during IETF in March 2015
- Coordination committees
  - CGP, JGP, and KGP met for 1.5 days in May with IP participation in some parts
  - A few more meetings may be needed to coordinate and conclude
  - Conclusion may be reached early next year
    - Complicated issues (as shown in the previous page)
    - KGP has had no experience on Han character domain names

# Framework of CJK LGR integration for Han characters



# Some of the Further Discussion Items

- Limiting the number of allocatable variant labels
  - Reduction of variant characters
  - Devising WLE with crafted definition of variant subtypes and rules
  - Investigation of the possibility of coordination between RootLGR and after-application evaluation
  - Investigation of the possibility for RootLGR to be empowered to accept 2 or more strings as applied-for strings
- Disposition of ‘distinct’ variants
  - Devising which are variants in some languages but are not variants in other languages, such as 机上 and 機上
  - Investigation of the possibility of coordination between RootLGR and after-application evaluation, by crafted definition of variant subtypes and outputs of RootLGR



# Cyrillic Generation Panel

Yuri Kargapolov, .УКР IDN ccTLD

Dusan Stojičević, .RS ccTLD / .СРБ IDN ccTLD

# What was discussed?

1. The “organizational” purpose of the first stage was formation of the document “*Proposal for the Generation Panel for the Cyrillic Script Label Generation Ruleset for the Root Zone*”
2. The “technical” purposes of the first stage was the established of the frameworks for future work of the Cyrillic Generation Panel. These frameworks should include multiplicity of the languages based on Cyrillic, lots of relevant Unicode code points, and the conditions under which the CGP can take a corrective decisions in future policy.

# What was discussed?

3. The Cyrillic GP took into account the following, conditionally speaking, "technical" features:

   - Panel's Diversity
   - Script for which the panel is to be established
   - Principal languages using the scripts that should include a number of languages according ISO 639-3
   - Geographic territories or countries with significant user communities for the script
   - The related scripts in Latin, Greek etc., and
   - Some features of subject, in particular, specific cases which are present in selected principal languages.

4. The Cyrillic GP took into account the following, conditionally speaking, "organizational" features:

   - Relevant experience and detail about organizations that represented in Panel
   - Relationship with Past Work or Working Groups within ICANN

# What have issues? (1)

## 1.2 Cyrillic scripts: Cyrillic - No. 220 Code Cyrl and Cyrillic (Old Church Slavonic variant) – No. 221 Code Cyrs

***Why two Scripts? Just question – what we will do if the Orthodox Church wants, in the future, to register own IDN gTLD which would be based on Cyrs? In doing so, Church in free, open, transparent way in many countries on all continents will gather in support one no less than 15 millions signatures that more the several times the population of some European countries.***

# What have issues? (1)

2. **We have a real problem in the existing “organizational” framework Cyrillic GP can not to cope alone. This is the problem of determining the set of “principal languages using the scripts that should include a number of languages according ISO 639-3”. We relied on the methodology for the determination of many ones through the use EGIDS level ( [https://www.ethnologue.com/about/language-status](https://www.ethnologue.com/about/language-status)).**

***Why? Because the levels of 4, 3 and 2 on the EGIDS scale for some languages that circulated in Russia require further study. The solution of this basic question depends to attraction to our team the specialists-linguists on Cyrillic languages.***

***We cannot recognize level of Rusin language because it has a different status and relevant EGIDS’ level in four countries (Serbia, Hungary, Slovakia and Ukraine) at least.***

***We found such professionals, but work on a volunteer basis, they will not.***

***We need help.***

# What have issues? (1)

3. We have highlighted a few particular cases that make us look on non-standard in some issues.

   - <u>Montenegrin case</u>. The national alphabet include as well as Cyrillic and Latin letters, but Latin Unicode code points (U+0179, U +017A, U+015A, U+015B). *Integration Panel withdrew thus issue because these code points are present in current version of MSR*.
   - <u>Ukrainian and Belarus case</u>. The national alphabets include apostrophe with function like usual letter.

# What have issues? (2)

- **the use of Uppercase and Lowercase Unicode code points in Cyrillic case. Despite at according RFC 5892 and IDNA 2008 Uppercase letters are disallowed, the users can enter letter of a domain name in any register, can receive email, for example, with “fishing” in any register; i.e. this issue connected with visualization.**

| TLD-string in Uppercase: The same letters which have “confusion variant” for users | | | TLD-string in Lowercase: The different letters which have not “confusion variant” for users | | |
|---|---|---|---|---|---|
| **.MAC** | = (in the perception of the user; visualization) | **.МАС** | **.mac** | != (in the perception of the user; visualization) | **.мас** |
| ***Latin*** | | ***Cyrillic*** | ***Latin*** | | ***Cyrillic*** |

# What have issues? (2)

4. We looked at only two multiplicity of Unicode code points which could lead to cross-script confusion variants – Greek and Cyrillic. Certainly, after report Armenian GP in Singapore, we had to draw attention to existence of a one more set. The Integration Panel did it in the comments.

| Armenian Script | Code Point | Cyrillic Script | Visual similarity |
|---|---|---|---|
| ա | U+0561 | ш (школа) | ա – Armenian<br>ш - Cyrillic |
| ո | U+0578 | п (пирог) | ո – Armenian<br>п - Cyrillic |
| օ | U+0585 | о (окно) | օ – Armenian<br>о - Cyrillic |
| պ | U+057A | щ (щенок) | պ- Armenian<br>щ - Cyrillic |

# What have issues? (3)

5. As noted by the Integration Panel is difficult to judge the adequacy of coverage of the Cyrillic GP. This is primarily due to the fact that the participants in the panel should cover not only the huge territory from the Balkans to the Pacific Ocean but over 108 languages.

**The Balkans and Eastern Europe are well represented in Cyrillic GP. Deficiency of the representation from Russia and Central Asia in all languages diversity. But if we found two specialists from Central Asia (now we have negotiation for inclusion in work of CGP), it is very hard to find the Chukchi, Tunguska, Kamchatka, Adyghe, Avar, Dargwa etc. IT-specialists who know the ICANN interests in this matter.**

***5 languages on EGIDS level 2 (Provincial)***

***2 languages on EGIDS level 3 (Wider communication)***

***11 languages on EGIDS level 4 (Educational)***

***We know such professionals in linguistic circles, but work on a volunteer basis, they will not.***

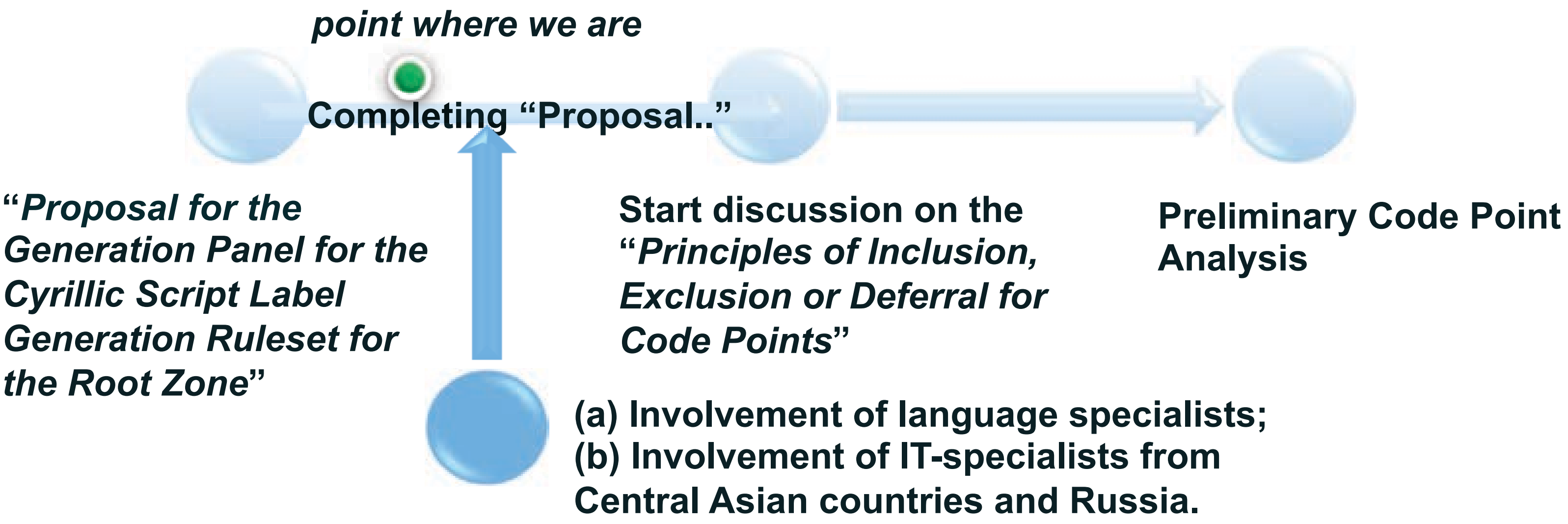
***We need help.***

# What next?

## 1.We agree with remark of Integration Panel:

*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).*

## 2.Next 3 steps



***point where we are***

**Completing "Proposal.."**

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

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

**Preliminary Code Point Analysis**

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

# Thanks!

# Engage with ICANN

ICANN

## Thank You and Questions

Reach us at: idntlds@icann.org
Email: engagement@icann.org
Website: icann.org

| | |
|---|---|
| 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 |