
ICANN
ICANN | 54
Dublin
18-22 OCTOBER 2015

IDN Program Update

Sarmad Hussain | IDN Program Sr. Manager | 21 October 2015

Overview of Session Presentations

- IDN Program Overview and Progress - Sarmad Hussain
- IDN LGR Toolset - Marc Blanchet
- Reference Second Level LGRs - Asmus Freytag
- Community Updates
  - Latin Generation Panel - Chris Dillon
  - Khmer Generation Panel - Rapid Sun
  - Thai Generation Panel - Wanawit Ahkuputra
  - CJK Generation Panel Coordination - Hiro Hotta

IDN Program Overview and Progress

Sarmad Hussain
IDN Program Senior Manager
ICANN

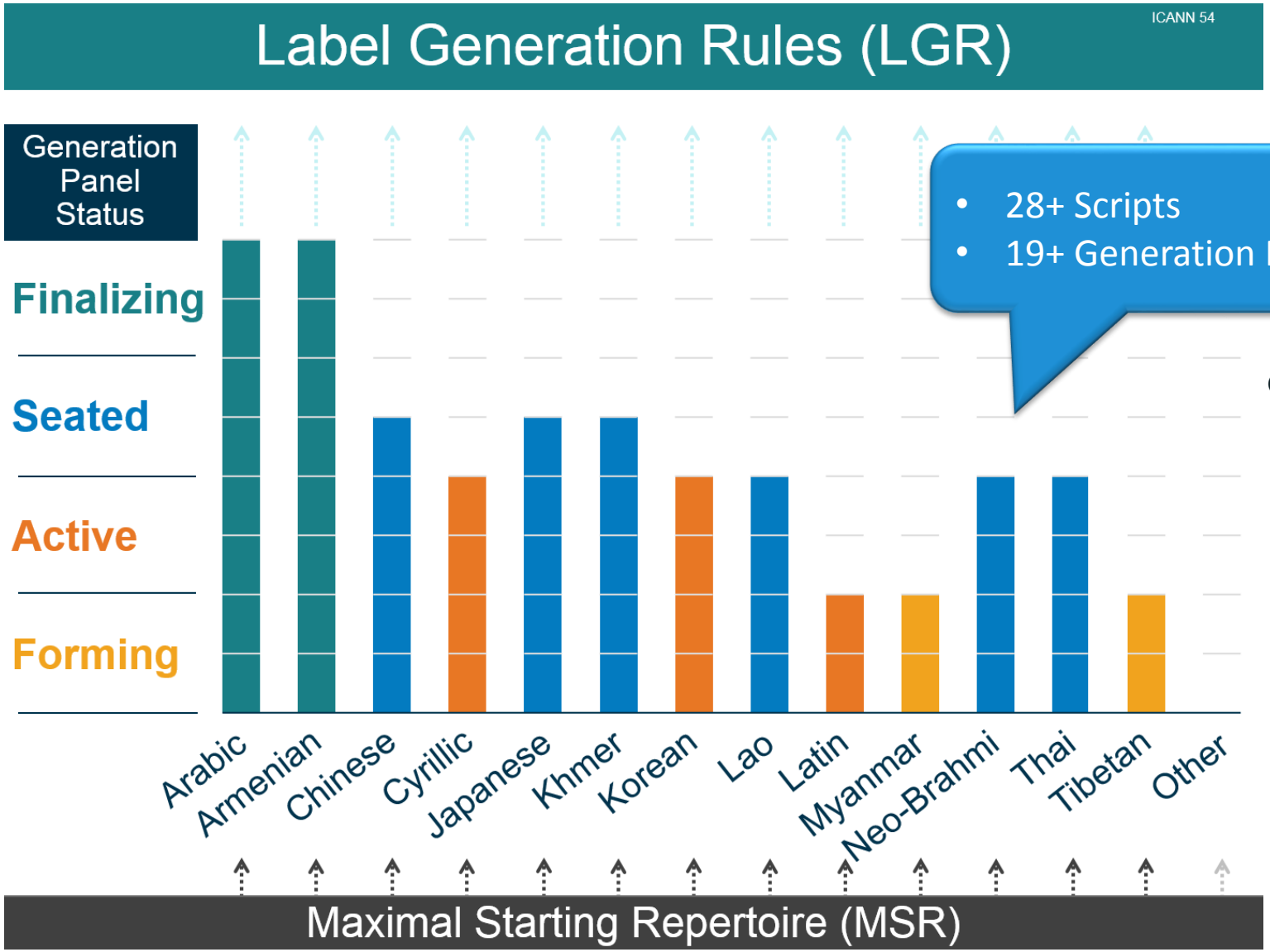
Overview of Presentation

- IDNs at Top Level
  - IDN TLD Program
    - Label Generation Ruleset (LGR)
    - LGR Toolset
    - IDN Variant Implementation
  - IDN ccTLD Fast Track Process Implementation
- IDNs at Second Level for gTLDs
  - IDN Implementation Guidelines
  - Reference LGR
- Community Outreach and Involvement

Root Zone Label Generation Rules (LGR)

- Generation Panel Support Documents
  - Guidelines released for GPs on 27 April 2015
  - Technical documents:
    - Variant Rules
    - Whole Label Evaluation (WLE) Rules
    - Representing Label Generation Rulesets using XML
  - Requirements for LGR Proposals
  - LGR Proposal Template
- Maximal Starting Repertoire (MSR)
  - MSR-2 released on 27 April 2015
  - Total 28 scripts
  - Total 33,490 code points shortlisted from 97,973 candidates
  - Based on Unicode 6.3
  - Upwardly compatible with MSR-1

Status of LGR Development



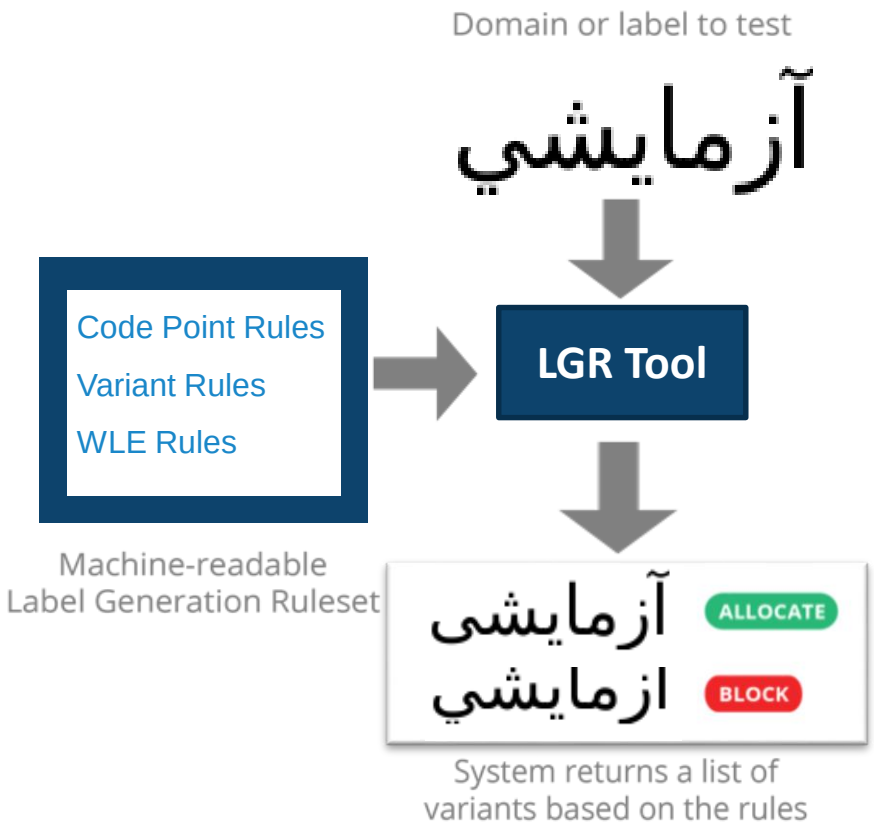
LGR Specification and Toolset

- LGR machine-readable specifications at:
  https://datatracker.ietf.org/doc/draft-davies-idntables

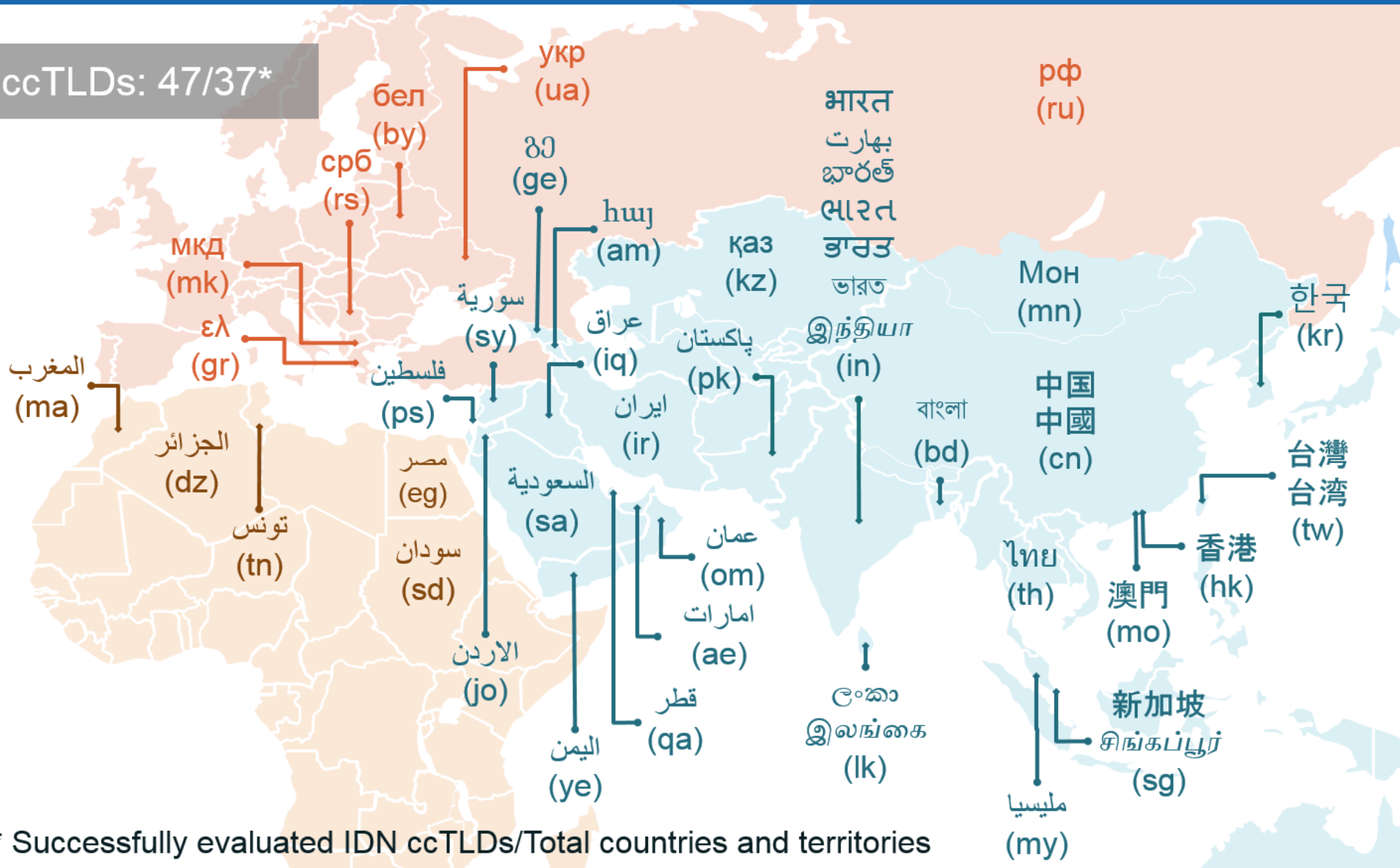
```
<xml>
        ...
        <char cp="06CC" >
                <var cp="0649" type="blocked" />
                <var cp="064A" type="allocatable" />
        </char>
        ...
</xml>
```

  - LAGER WG at IETF

- Toolset tentative timeline:
  - Create LGR - available
  - Use LGR – 12/15
  - Manage LGRs – 3/16
  - Open source



IDN Country Code Top-Level Domains



* Successfully evaluated IDN ccTLDs/Total countries and territories

IDN ccTLD Fast Track Process

- Completed 5+ years of operation
  - Requests in 18 scripts for 27 languages
  - 37 IDN ccTLDs delegated in the Root Zone representing 29 countries/territories

- Currently under annual review
  - Public comment announced on 15 Jan. 2015
    - Second similarity review and process
  - Public comment closed on 17 March 2015
  - Board resolution on string similarity review on 25 June 2015
  - ccNSO developing EPSRP working group

Reference Second Level LGRs

- Language tables submitted by new gTLDs intending to offer IDNs
- Reference being developed for facilitation and consistency in
  - Pre-Delegation Testing (PDT)
  - Registry Service Evaluation Process (RSEP)
- Registries may submit different tables with supporting documentation

- Current Status
  - ✓ Guidelines and documentation of authoritative sources
  - Public comments
  - IDN table in LGR format
  - Expert review: linguistic - security and stability
  - Public comments
  - Publication after incorporating public feedback

  - **Batch 1:** *Japanese, Korean, Chinese, Danish, Norwegian, Latvian, Lithuanian, Russian, Arabic, Ukrainian, Belarusian, Bulgarian, Macedonian, Bosnian (in Cyrillic and Latin scripts), Serbian, Hebrew*
  - **Batch 2:** *English, Spanish, French, German, Portuguese, Polish, Swedish, Italian, Hungarian, Icelandic, Finnish, Montenegrin*

IDN Implementation Guidelines

- Background and motivation
  - To promote IDN registration policies and practices and to minimize consumer risk and confusion
  - Last updated in 2011; GNSO community requested for review

- Current status
  - Call for Community Experts to Review the IDN Implementation Guidelines on 20 July 2015
  - WG formed with experts from ALAC(2), SSAC(1), gNSO(6) and ccNSO(2)

Communication and Outreach Efforts

- Updated IDN web pages at icann.org/idn
- IDN Program sessions at ICANN meetings
- IDN Program updates to SOs/ACs at ICANN meetings
- Presentations
  - Support IDN related outreach (APrIGF, ArmenianIGF, TLDCON)
  - Direct outreach (Thailand, Pakistan)
- Blogs
  - Linguistic Diversity in the Internet Root: The Case of the Arabic Script and Jawi – Rinalia Abdul Rahim
  - Collaborating towards a truly multilingual Internet
- ICANN Community Wiki LGR Project Pages
- IDN mailing lists
  - {vip, lgr, ArabicGP, ArmenianGP, ChineseGP, …}@icann.org

Useful Links for IDN Program

- IDN Program: http://icann.org/idn
- For any queries regarding the IDN Program, please email: IDNProgram@icann.org

- To join a Generation Panel for your language, submit CV and statement of interest at: idntlds@icann.org;
- Call for Generation Panels: http://www.icann.org/en/news/announcements/announcement-11jul13-en.htm
- LGR Document Repository: https://community.icann.org/display/croscomlgrprocedure/Document+Repository

IDN LGR Toolset

Marc Blanchet

Agenda

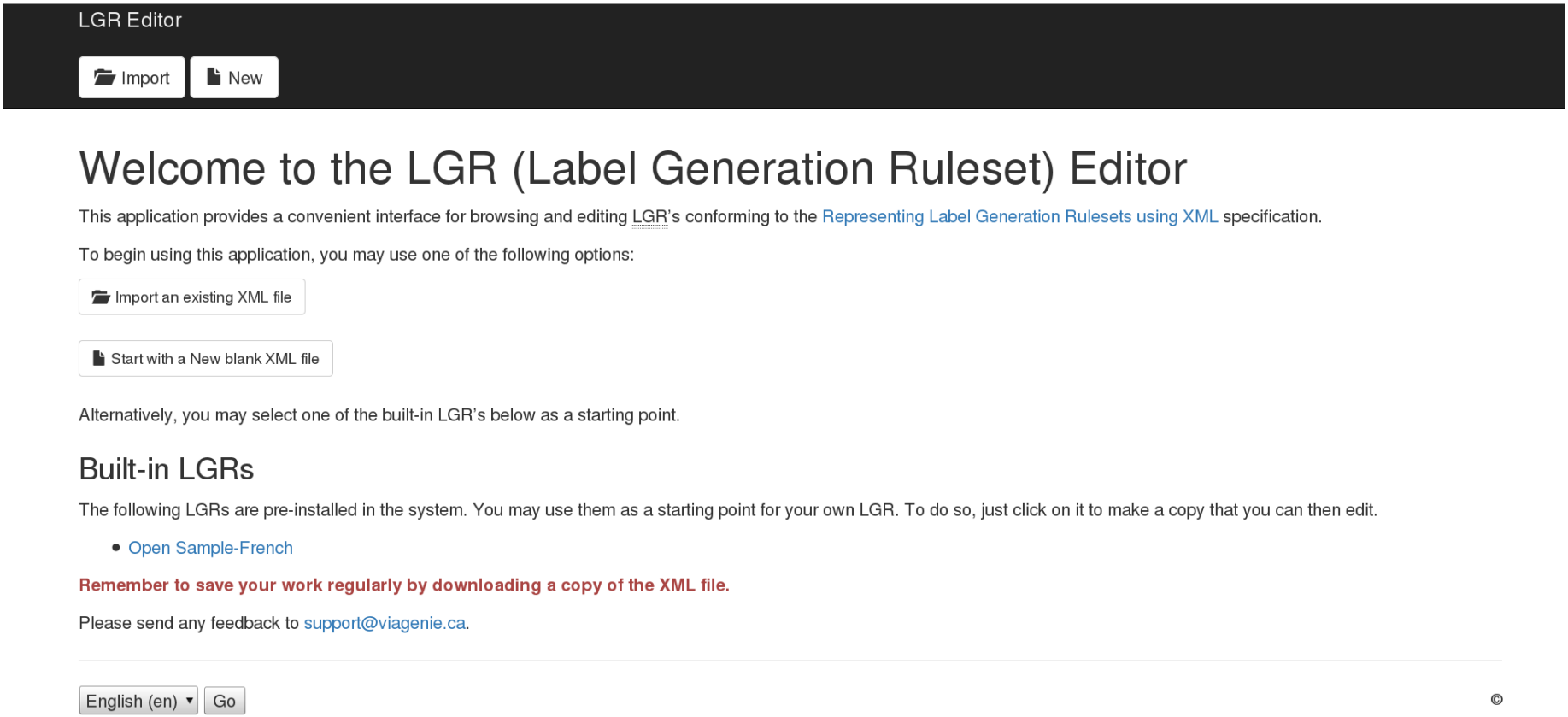
1. Background
2. Walkthrough Project Plan and Timeline
3. Current Status
4. Conclusion

Background

- Tool to help LGR designers create their LGR
  - Web front-end
  - Open source
  - Define and manage variants
  - Validations
  - Labels to test against, …
  - LGR XML format can be complicated for some use cases and is cumbersome for non-XML savvy people
- 3 phases:
  - LGR Edition tool. Released August 2015
  - current -> Validate labels, generate variants. To be Released November 2015
  - LGR management tool: merge, diff, etc.

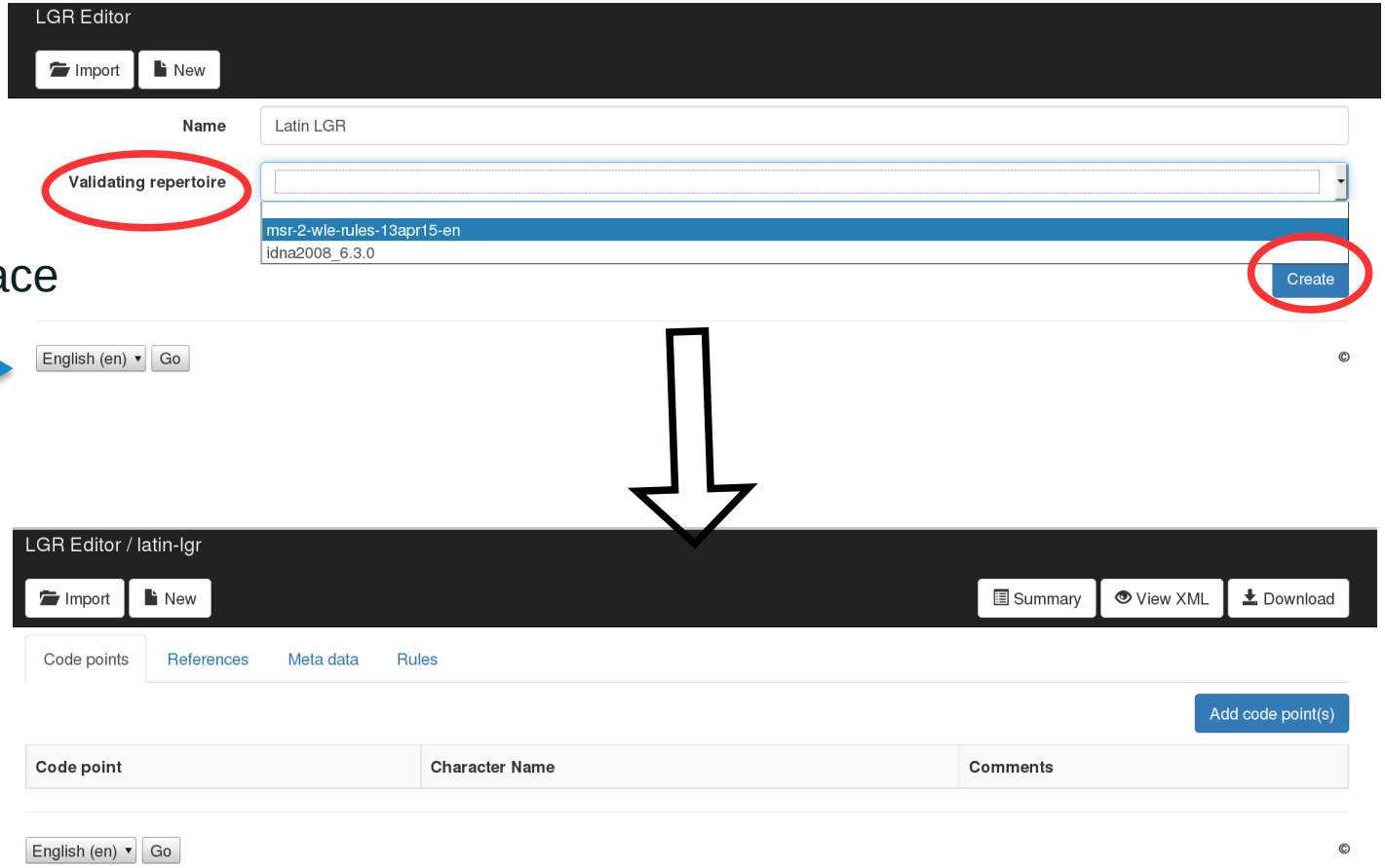
Example: Walkthrough with a French LGR

Welcome Screen

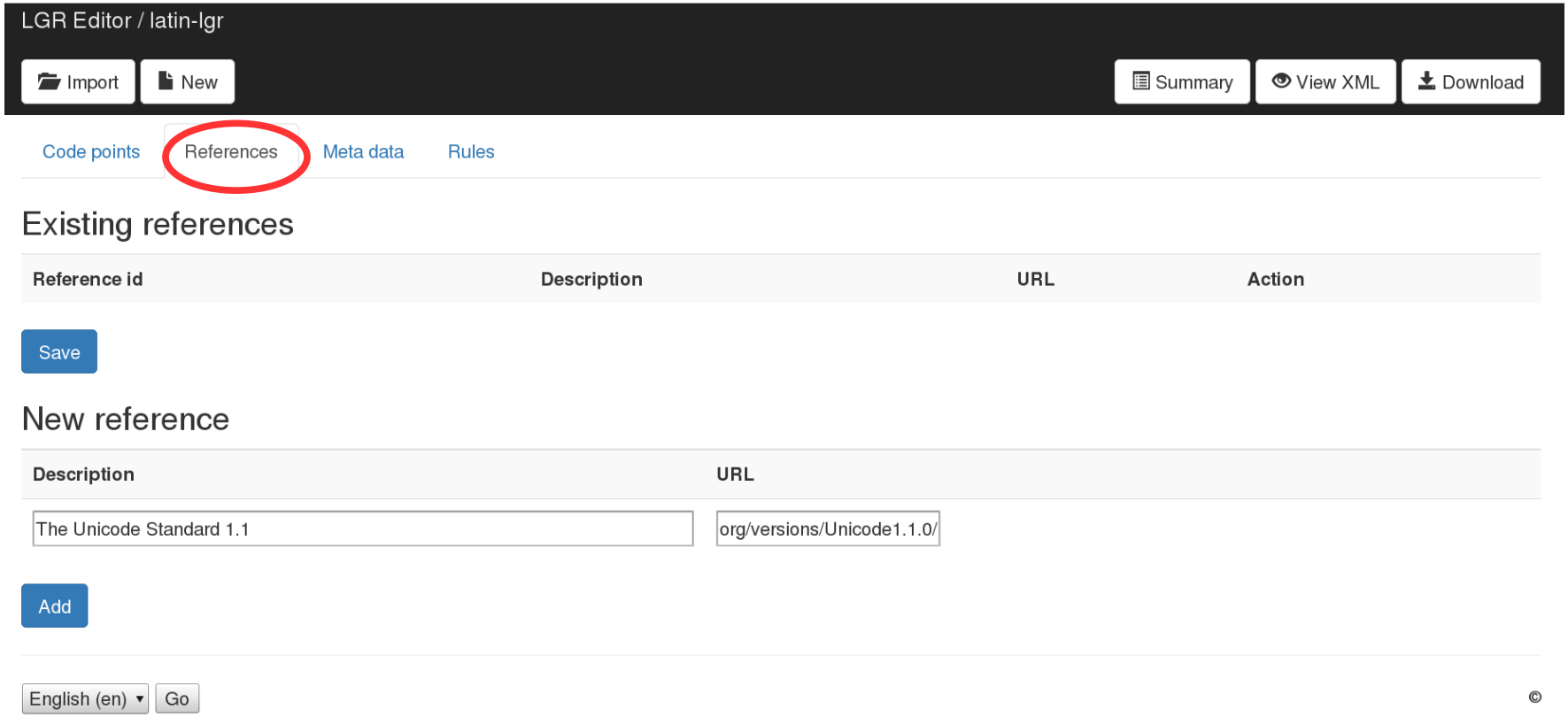


Create New LGR

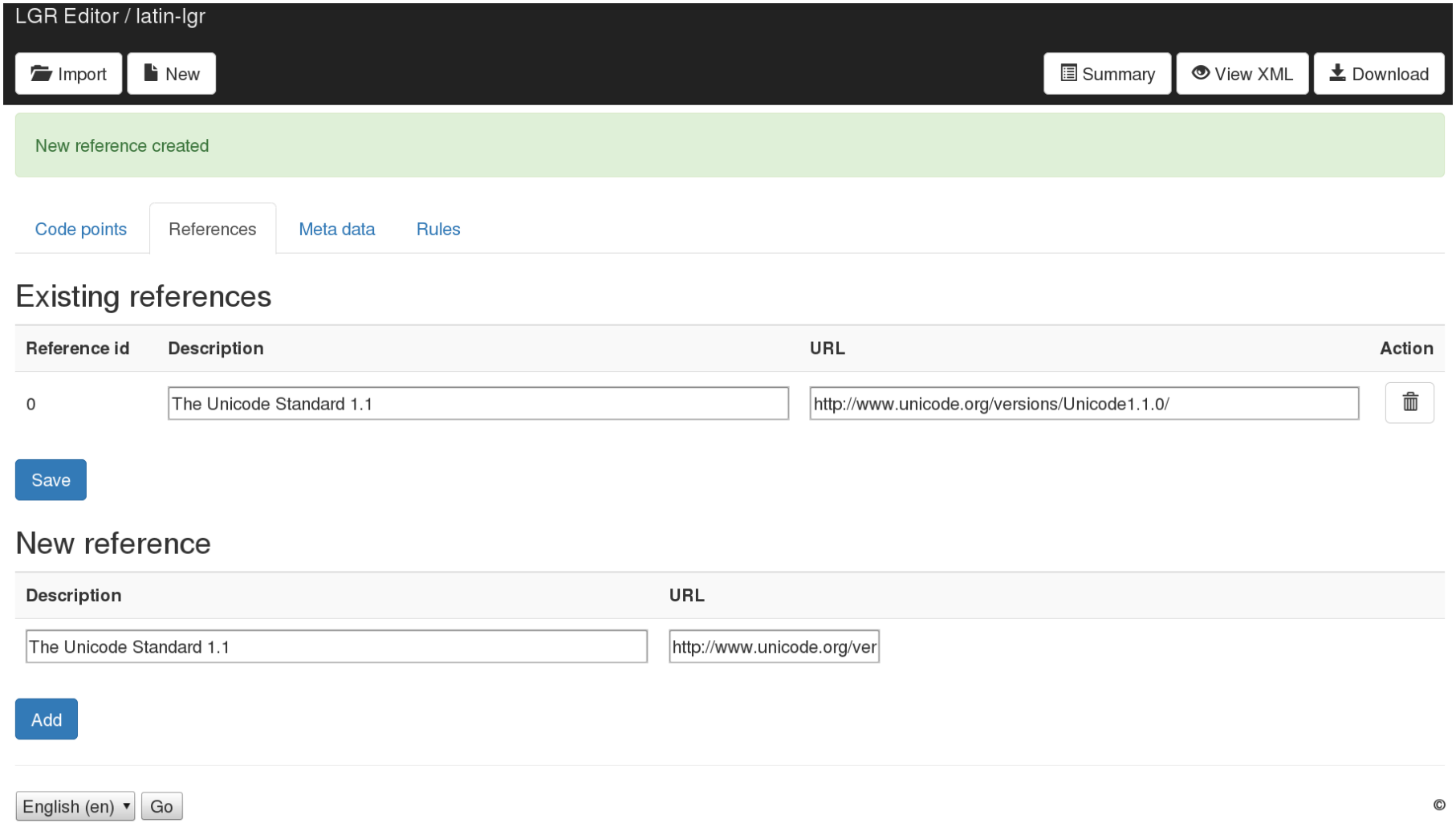
Tool Interface Language



References

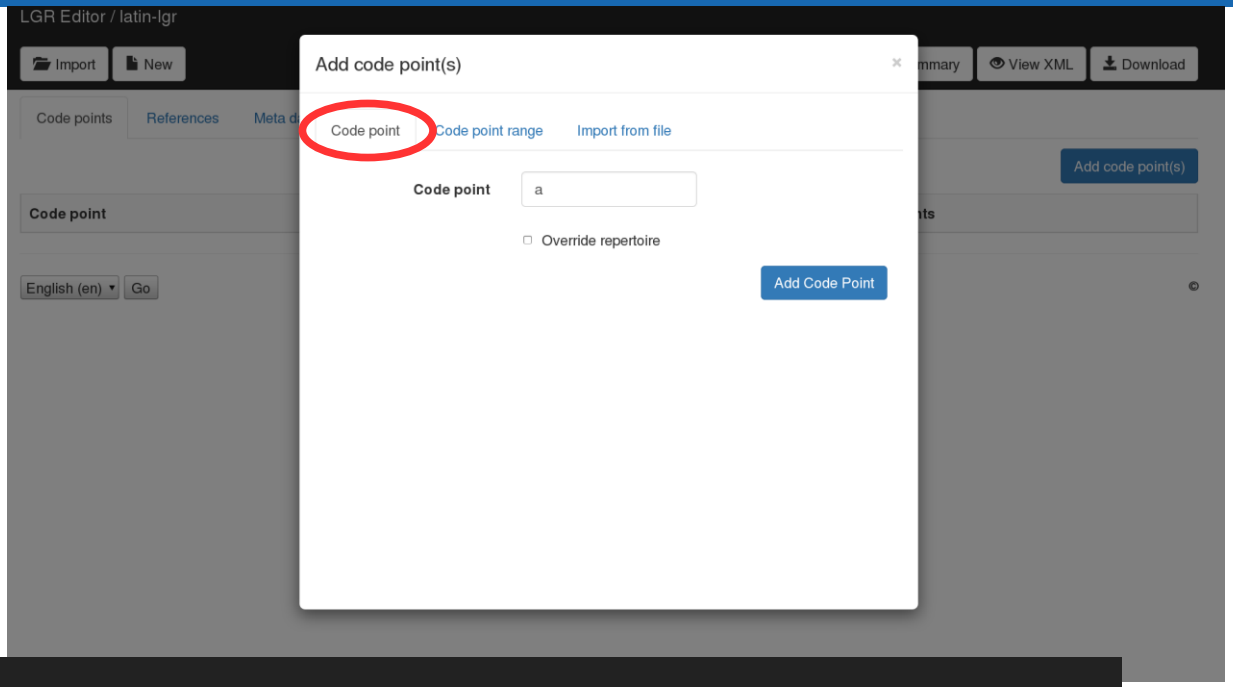

LGR Editor / latin-lgr
Import
New
Summary
View XML
Download
Code points
References
Meta data
Rules
Existing references
Reference id
Description
URL
Action
Save
New reference
Description
URL
The Unicode Standard 1.1
org/versions/Unicode1.1.0/
Add
English (en)
Go
©

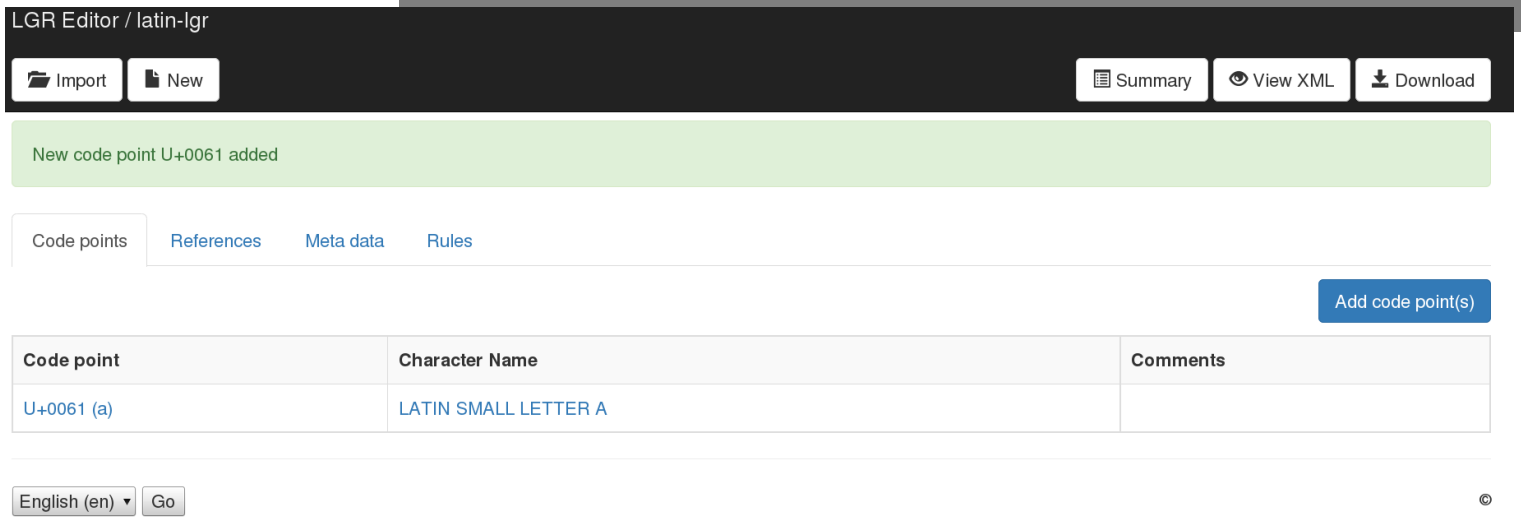
References (cont.)



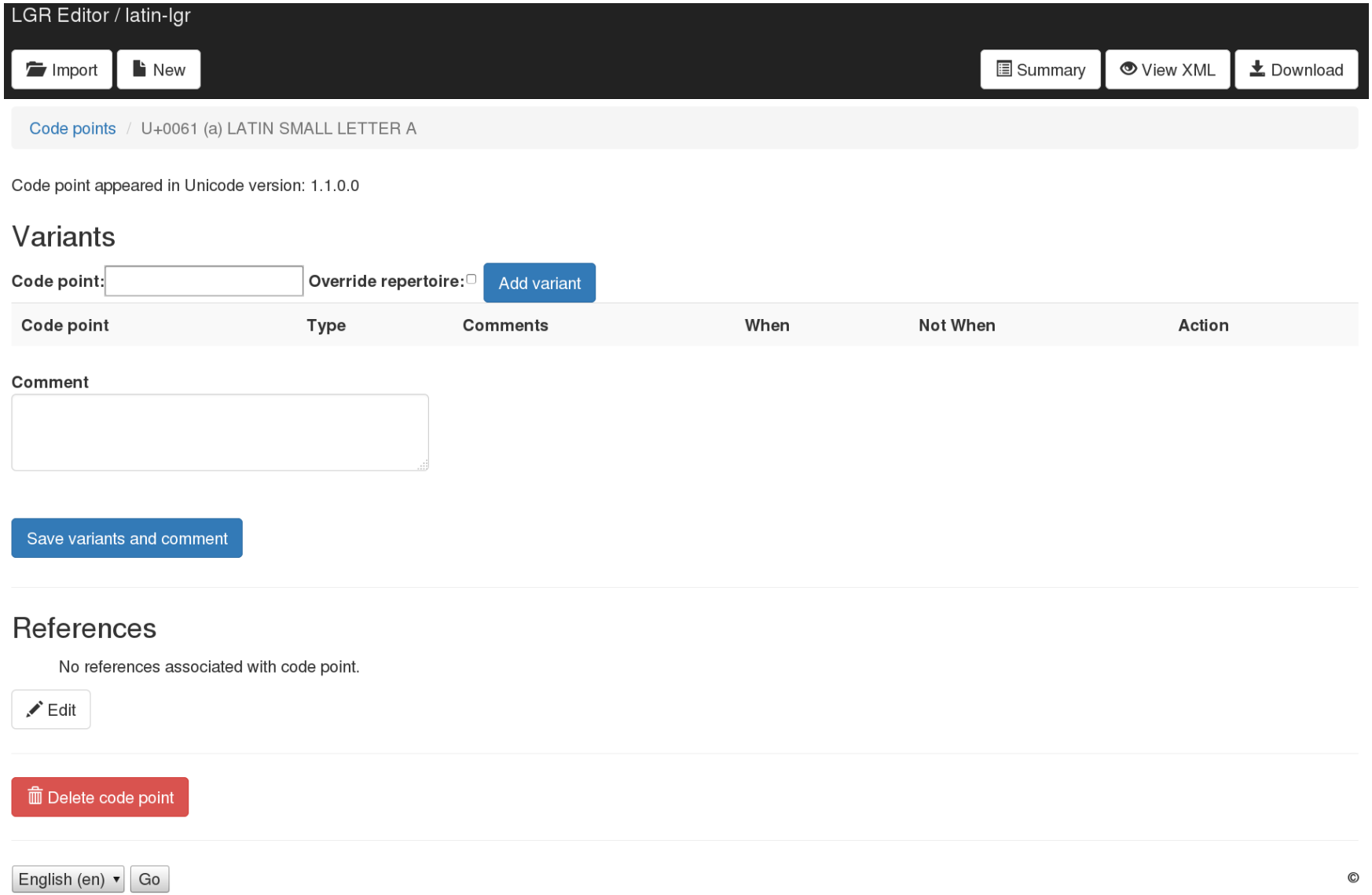
Code Point

Code point input can be:
- Direct
- U+ Unicode notation
- Sequence
  (blank space between)

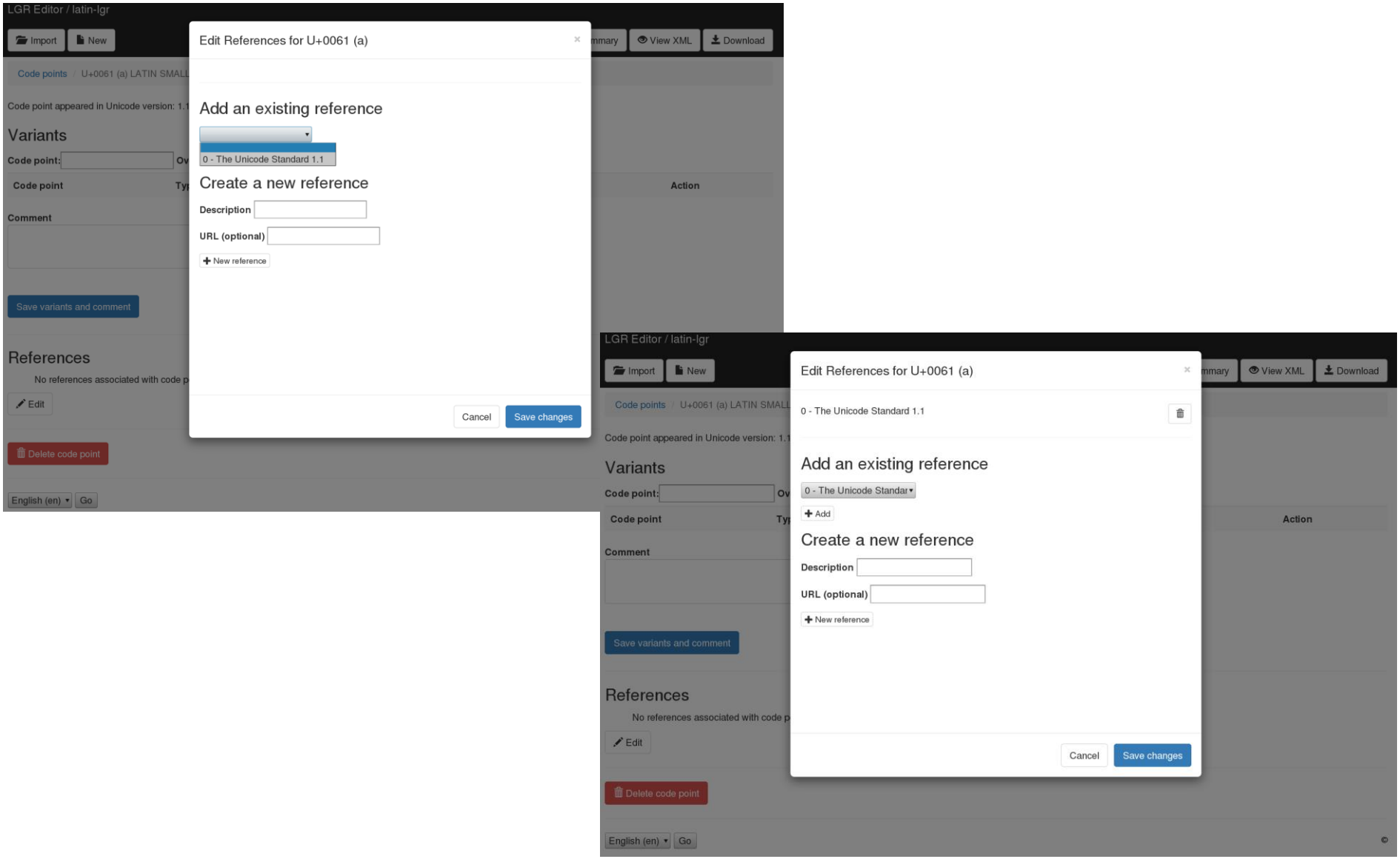




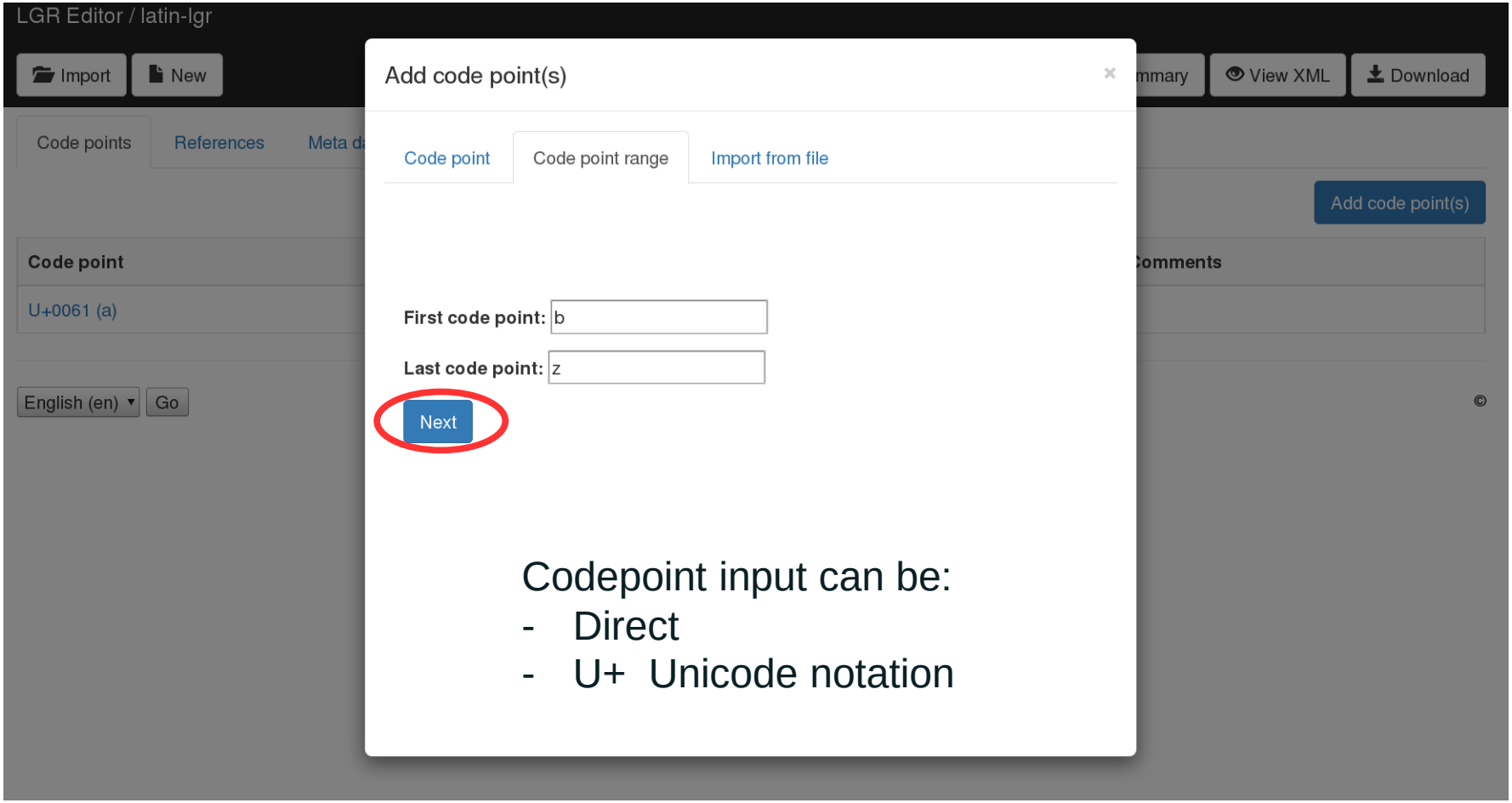
Editing Code Point



Editing Code Point - References

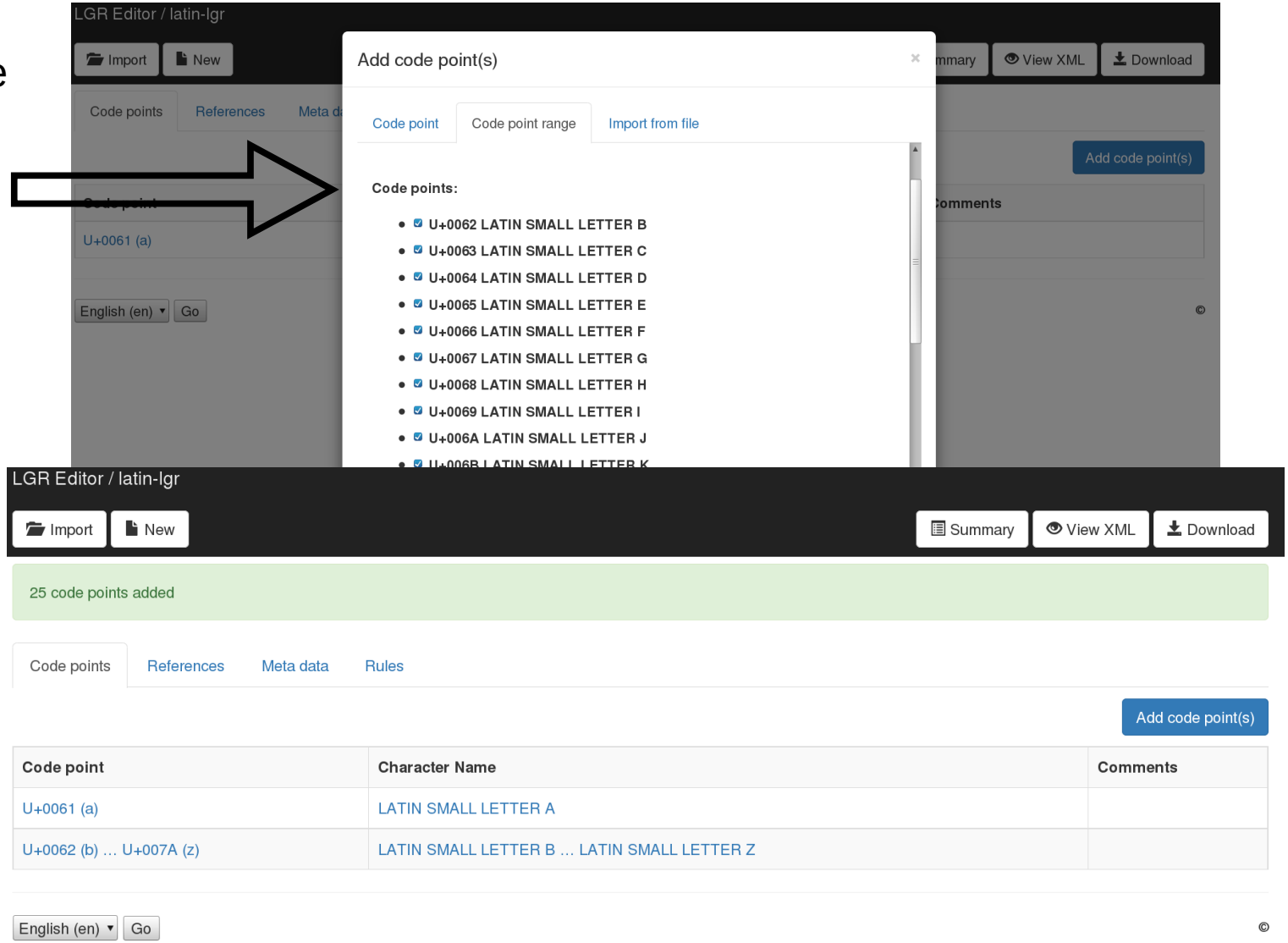


Range


LGR Editor / latin-lgr
Import
New
View XML
Download
Code points
References
Add code point(s)
Code point
Code point range
Import from file
First code point: b
Last code point: z
Next
Code point
U+0061 (a)
Comments
English (en)
Go
Codepoint input can be:
- Direct
- U+ Unicode notation

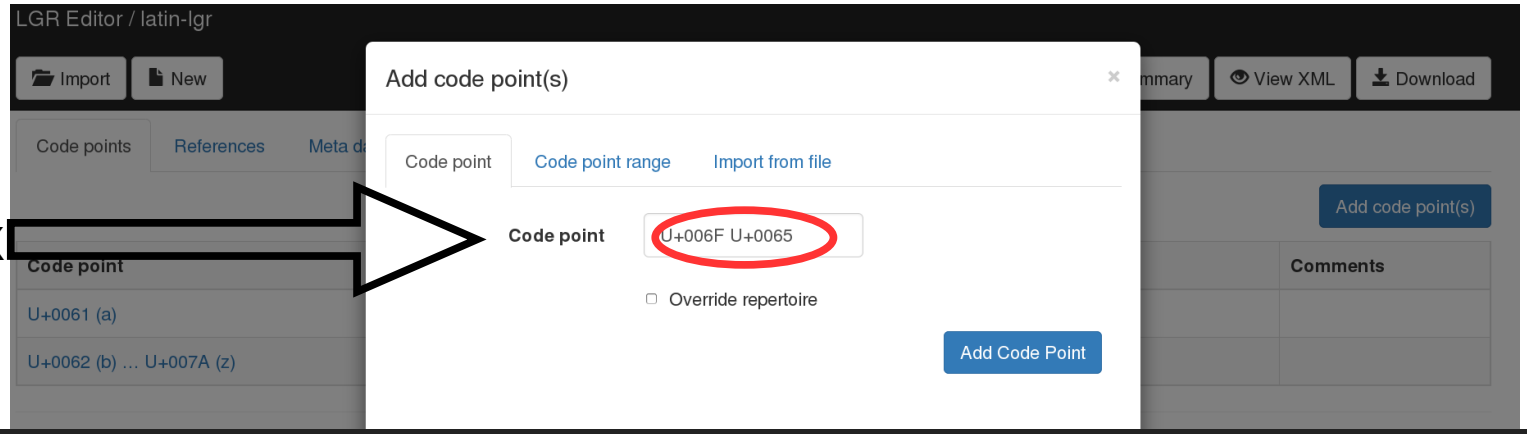
Range (cont.)

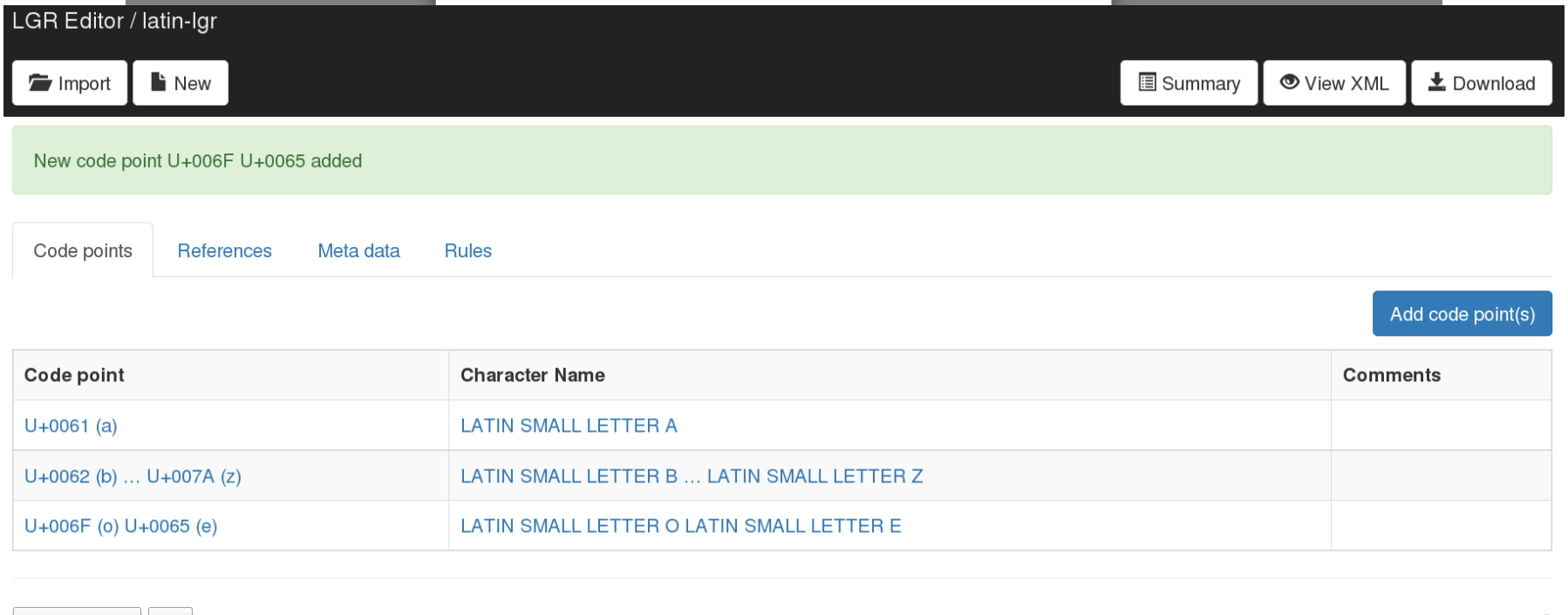
Choose code points from range



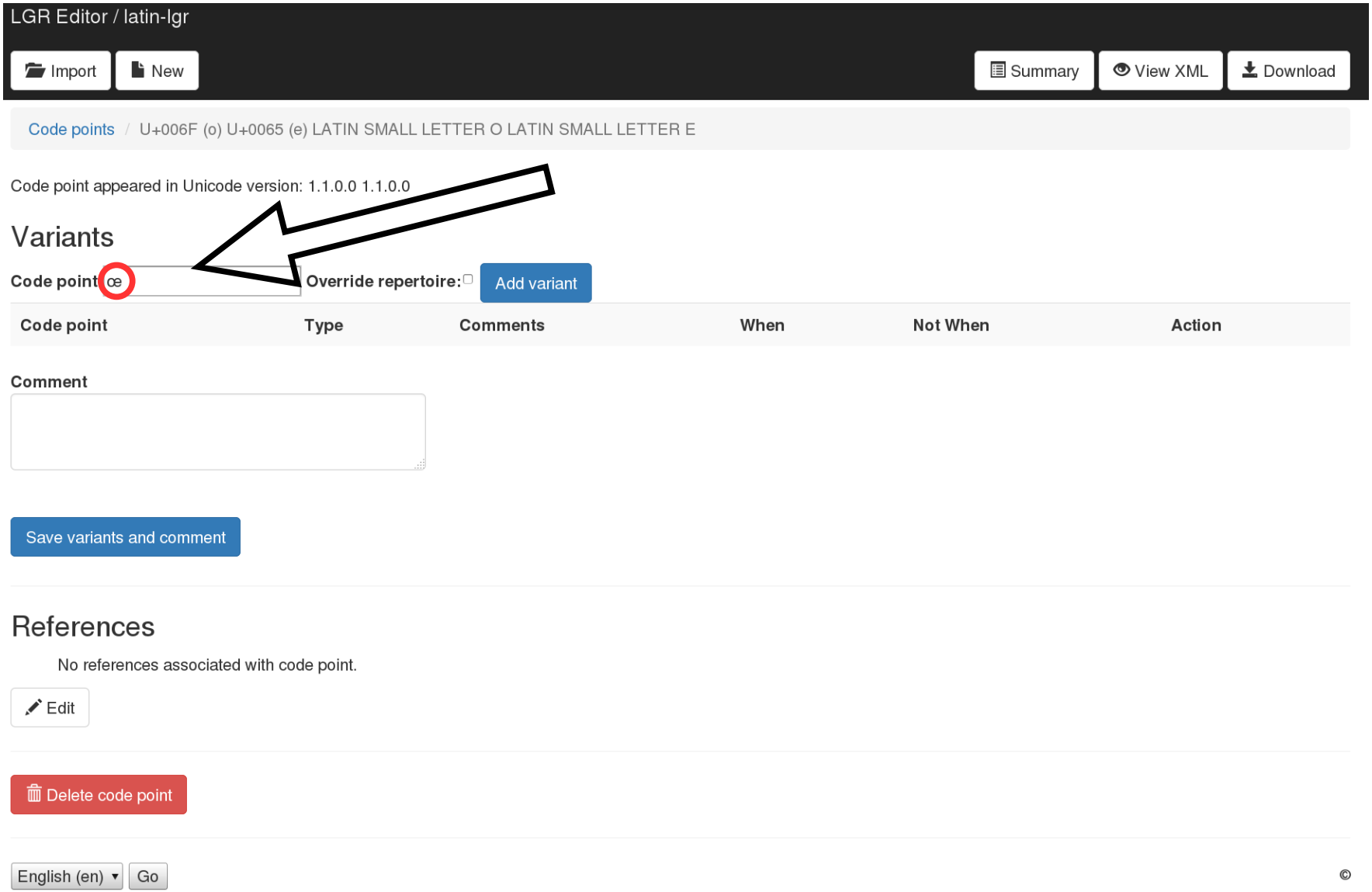
Code Point Sequence

U+XXXX format

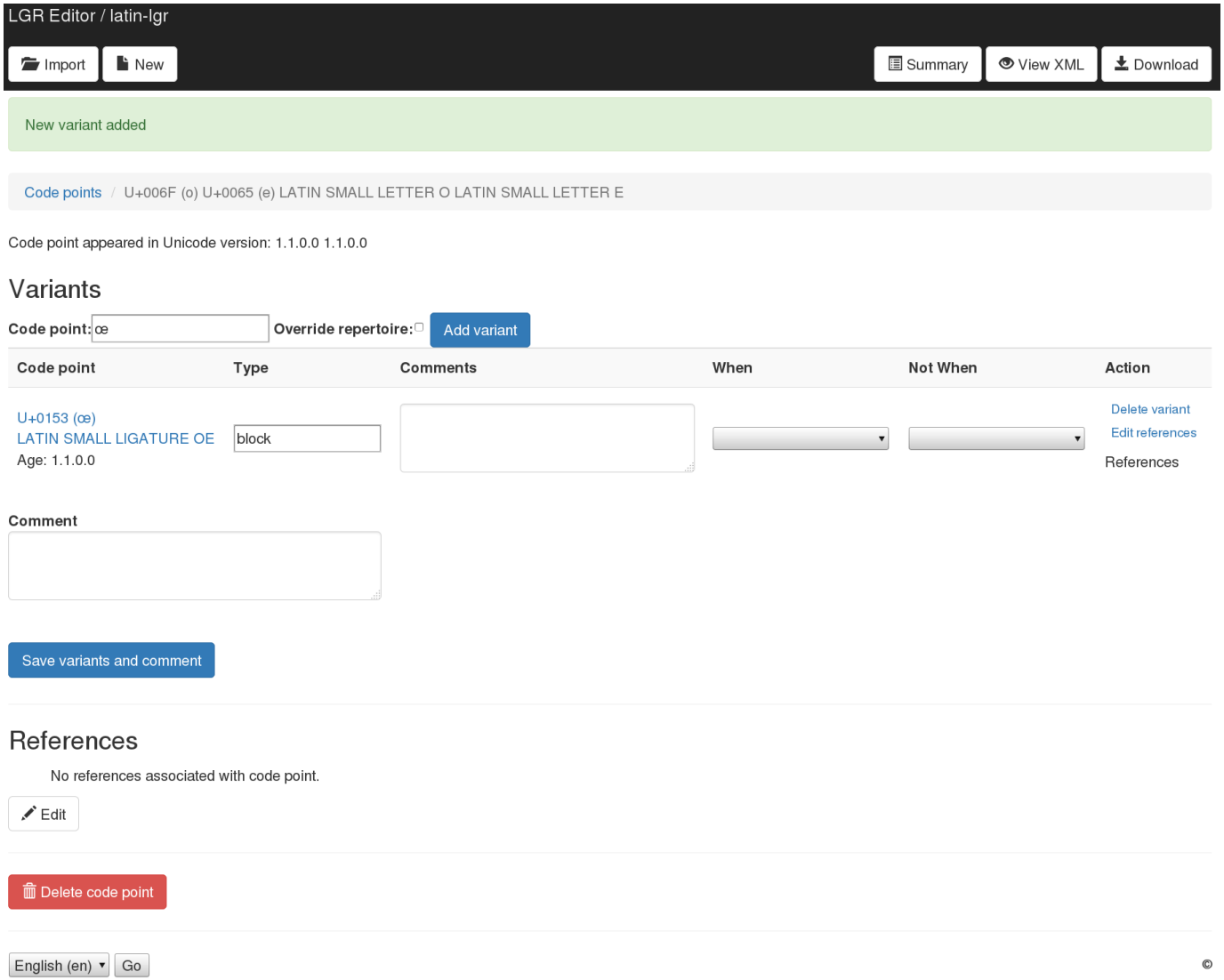




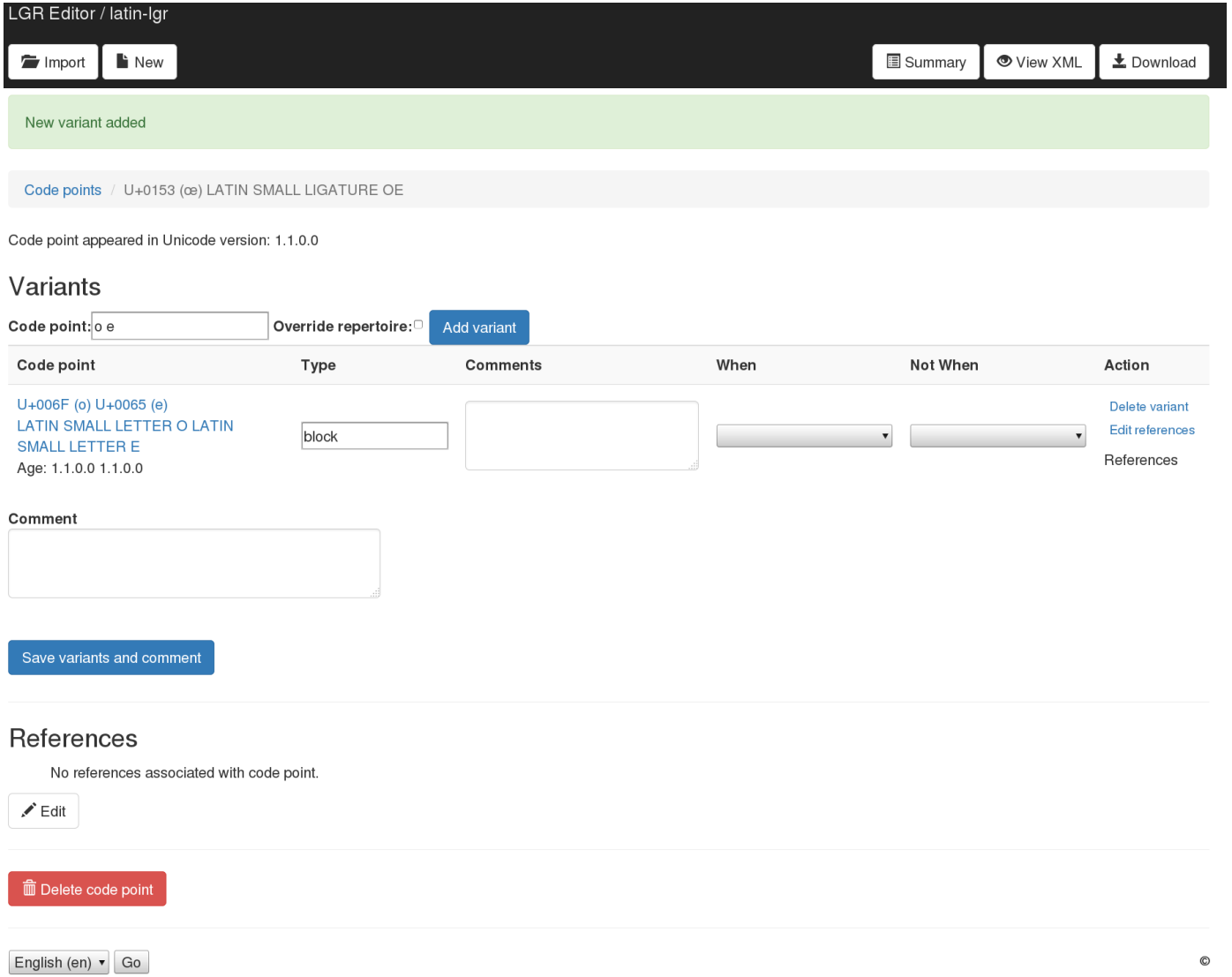
Variant


LGR Editor / latin-lgr
Import
New
Summary
View XML
Download
Code points / U+006F (o) U+0065 (e) LATIN SMALL LETTER O LATIN SMALL LETTER E
Code point appeared in Unicode version: 1.1.0.0 1.1.0.0
Variants
Code point œ
Override repertoire:
Add variant
Code point
Type
Comments
When
Not When
Action
Comment
Save variants and comment
References
No references associated with code point.
Edit
Delete code point
English (en)
Go
©

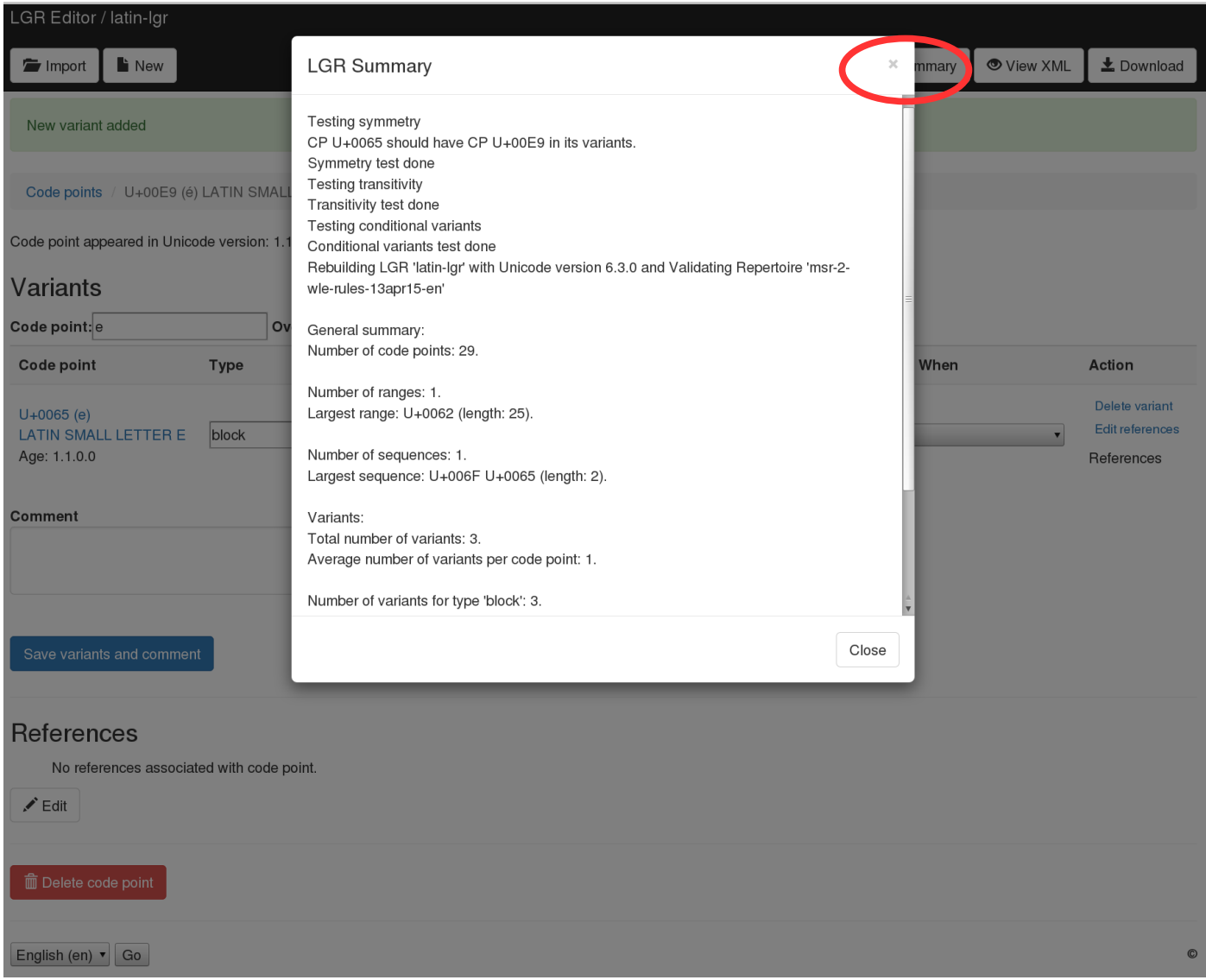
Variant (cont.)



Variant - Symmetry



- Validate the LGR
- Provides various statistics
- Can be compute intensive, depending on the LGR



Current Status

- Available at: http://lgr-demo.viagenie.ca
  - To get access, mail: marc.blanchet@viagenie.ca
  - Already many users
  - Updated as new code is stable enough
- Phase 2 code almost done; on target for November 2015 release
- Any questions? Mail: marc.blanchet@viagenie.ca

Reference Second Level LGR

Asmus Freytag

Agenda

- Guidelines for language-based Reference LGRs for the second level
  - Purpose
  - Existing Work
  - Review Process
  - Deliverables

Purpose

- Describe the development process for language-based reference LGRs for the second level

- Clarification of terms:
  - Language within the context of writing systems and script
  - Some languages use a choice of scripts, others need several scripts

- Target repertoire
  - Essential, common use and extended subset

- Sources
  - What constitutes an authoritative source?
  - Few sources provide data specific to IDN labels
    - Lack of authoritative sources for variants, WLE rules, etc.
    - Necessity to apply judgment

Starting Point: Language Tables (.SE)

- .SE (The Internet Infrastructure Foundation) created a set of 29 Language Tables and a guideline document
- Intended as starting point
- Tables are expressed in legacy text format
  - Not adequate for expressing WLE and variant rules
  - In need of more authoritative sources
    - Omniglot and Wikipedia were the only sources used in most cases

Documenting the Required Repertoire

- Challenges in verifying and documenting the repertoire
  - Not all languages have institutional authorities
  - Only some have de-facto authorities
- Essential subset:
  - Sources generally agree; using “better” sources does not improve the results
- Common use subset:
  - Actual set to cover usual spelling of words in the language
  - “Authoritative” sources can be incomplete or open-ended
    - For example, German “Rat für Rechtschreibung” lists just a few example code points (“such as…”)
- To get useful results for IDN, need to consult additional sources

Existing Work (CLDR and IDN ccTLD)

- Common Locale Data Repository (CLDR)
  - Maintained by the Unicode Consortium,
  - Based on input from local experts
  - For each language, contains a specification of:
    - Core set that captures the **essential set** of code points
    - Auxiliary set that captures the **maximal set** of code points
- CLDR data collection is an open process supported by multiple vendors and the result is widely implemented in the industry
- IDN ccTLD language tables also provide useful input
  - When they involve languages native to the country or territory
    - Examples: Japan, countries using the Cyrillic scripts

Review by Linguistic Experts

- Review by linguistic experts will focus on these main points:
  - Does the LGR omit required code points, variants, rules?
  - Does the LGR omit desirable code points, etc.?
  - Does the LGR include unnecessary code points, etc.?
  - Does the LGR include undesirable code points, etc.?
  - Is the documentation relevant and authoritative?
    - Would better sources lead to different outcomes?
  - Does the XML accurately describe the LGR?
  - Are labels outside the strictest subset adequately supported?
  - Does any code point, variant or rule cause issues for the LGR?

Review for DNS Stability and Security

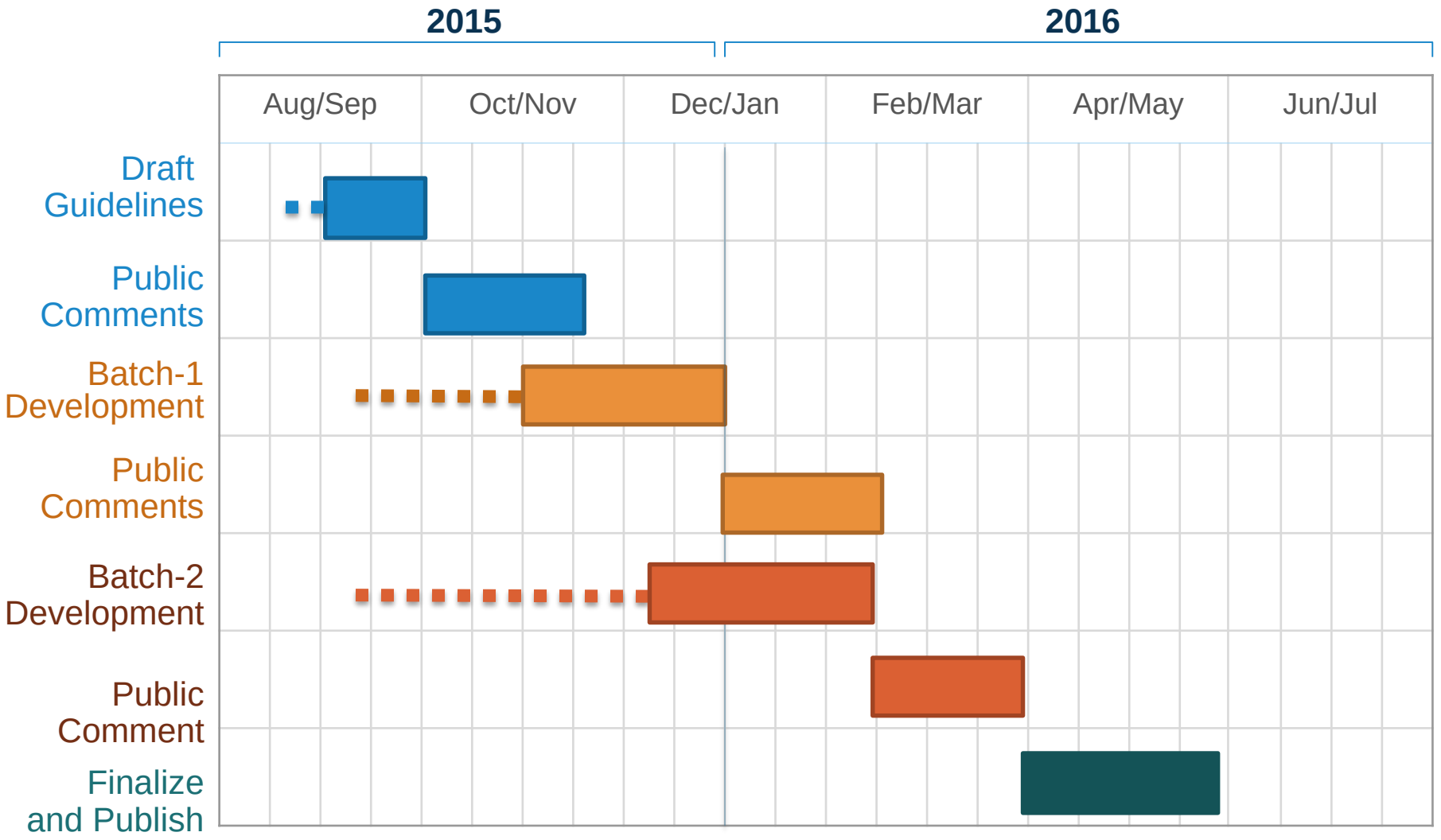
- Expert reviewer separate from linguistic reviewer

- Questions to be considered:
  - Does the repertoire allow undesirable script mixing?
  - Are CONTEXTO/CONTEXTJ code points allowed?
    - Is that choice justified and context rules provided?
  - Are combining marks limited to fixed sequences?
    - If not, are they properly restricted via rules?
  - Are there stability/security concerns for code points?
  - Are there stability/security concerns for variants?
    - Any additional variants required for security?
  - Are there stability/security concerns for rules?
    - Any additional rules required for security?
  - Does the LGR address any issues related to the principles from RFC 6912?

Deliverables and Follow Up

- Planned deliverables for each language-based LGR
  - XML file
  - Descriptive document
  - Expert reports will be attached
- Submission for public comments
  - Any changes required by public comments will be made
  - Experts will re-review as needed and reports will be updated
- Final expert reports will be attached to final LGR documentation

Project Timeline



Thank You!

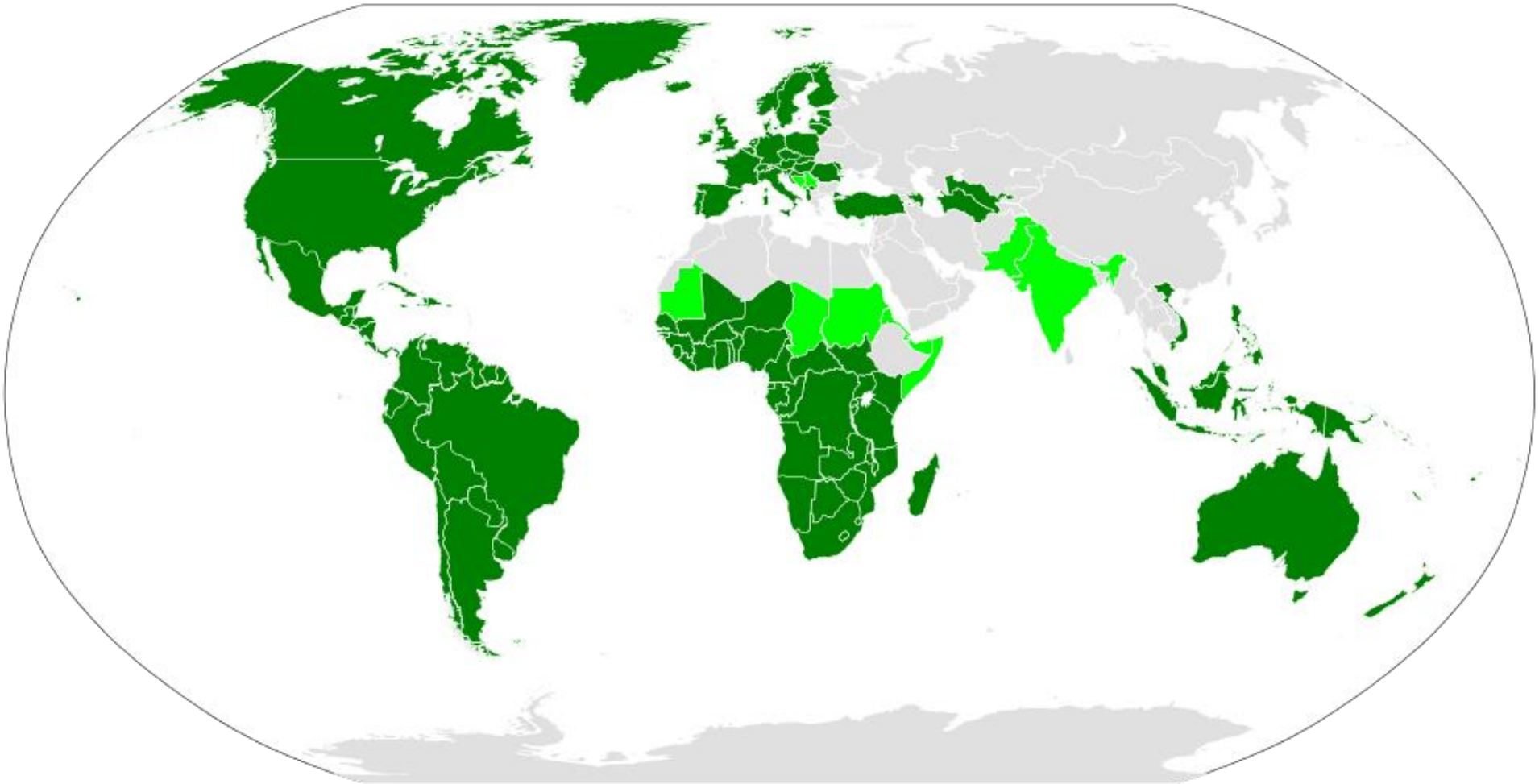
Community Update:
Latin Generation Panel

Chris Dillion
Co-chair, Latin GP

Agenda

1. Potential scope of the Latin Generation Panel
2. Case Study
3. Members of the Latin Generation Panel
4. Additional Expertise Needed
5. What Next?
6. Questions and Contact Details

Distribution of the Latin Script



See also: www.omniglot.com/writing/langalph.htm#latin

Light green: countries where Latin co-exists with other scripts.

Map by Canuckguy.

Example: Latin Script Use in Africa Today

- Today, the Latin script is the writing system in widest use in Africa
  - It is estimated that over 500 out of the 2000 languages spoken in Africa today have orthographies (Bendor-Samuel 1996: 689), with the vast majority being Latin script-based
- The Latin script has been significantly extended or modified to represent African languages:
  - Frequently, supra-segmental features such as tone were encoded using super-and subscripted graph(eme)s, such as accent marks
  - Next to entirely new letters, di-, tri- and quadrigraphs, for example, are often much used to represent single phonological units
  - A number of code-points are already excluded by the “letter principle” in the MSR, as well as IDNA 2008

    Meikal Mumin
  - The situation is similar for indigenous languages in the Americas

Example: Romanizations of Other Scripts

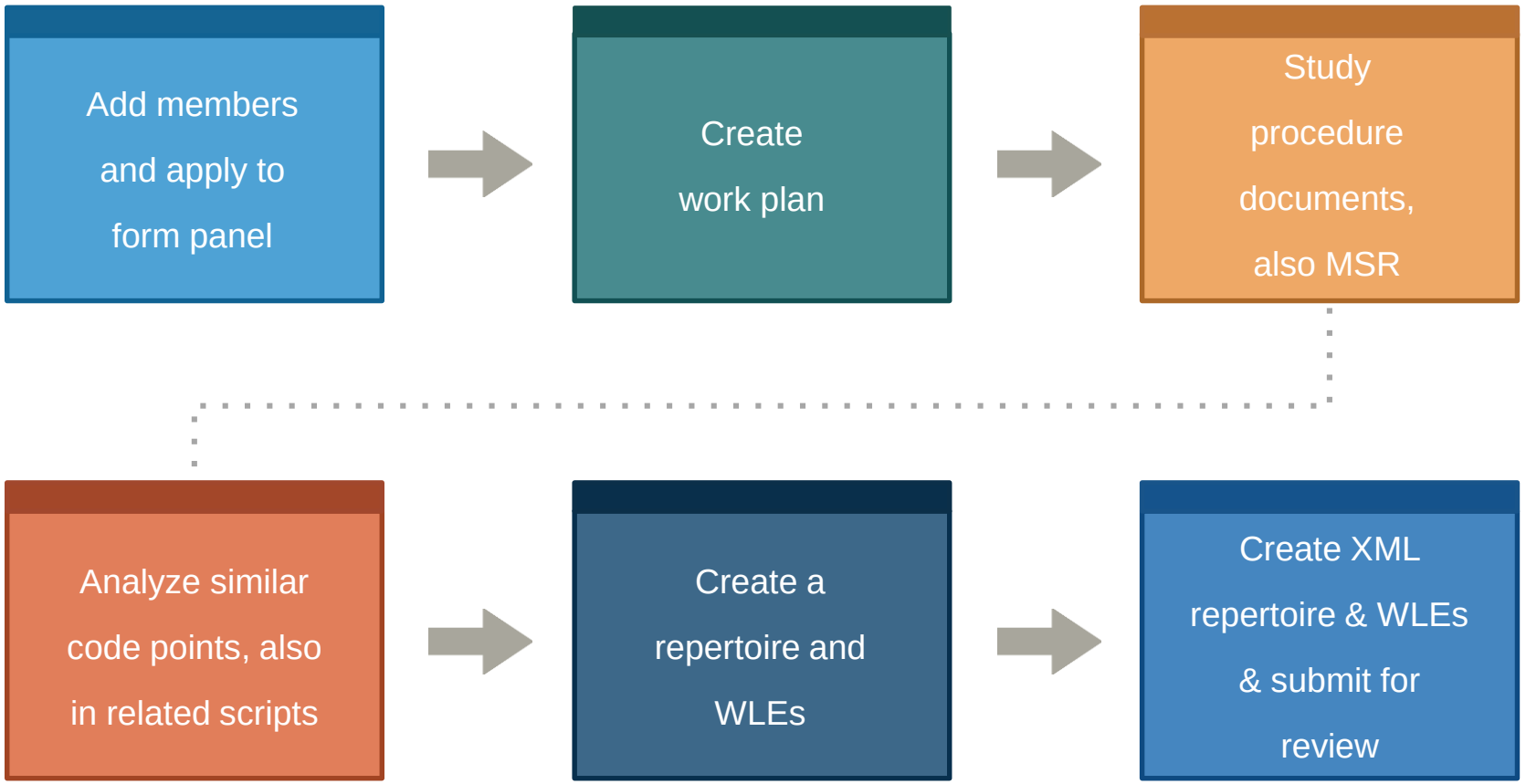
- The Latin script is used to Romanize other scripts:
  - Standardized Romanizations such as Pinyin for Chinese:
    Hěn gāoxìng jiàndào nǐ
  - Informal Romanizations such as Arabic chat:
    ana raye7 el gam3a el sa3a 3 el 3asr

Current Membership of Latin Generation Panel

Name	Country	Language Expertise
Eric Brunner-Williams	US	English
Chris Dillon	UK	English, Japanese, German, Spanish, Korean, etc.
Hazem Hezzah	Egypt	Arabic, German
Paul Hoffman	US	English
Matthias Brenzliger	South Africa	
Tunde Adegbola	Nigeria	
Danko Jevtovic	Serbia	Serbian, English
Jean-Jacques Subrenat	France	French, English
Yashar Hajiyev	Azerbaijan	Azerbaijani, English

Name	Country	Language Expertise
Mirjana Tasić	Serbia	Serbian, English
Oscar Gabriel Ledesma Piñeiro	Argentina	Spanish, English
Daniel Omondi	Kenya	
Gideon Kiprono Rop	Kenya	
Tarik Merghani	Sudan	
Meikal Mumin	Germany	German, English, Italian, French, Arabic, etc., Latin for African languages
Tarkan Doruk	UAE	Turkish
Mert Saka	Turkey	Turkish
Ahmed Bakht Masood	Pakistan	Urdu, English

What Next?


Add members and apply to form panel
Create work plan
Study procedure documents, also MSR
Analyze similar code points, also in related scripts
Create a repertoire and WLEs
Create XML repertoire & WLEs & submit for review

Community Update:
Khmer Generation Panel

Rapid Sun
Secretary, Khmer GP

Agenda

- Introduction to Khmer Language
- Introduction to Khmer Script
- Current membership of Khmer GP
- Current progress – code point repertoire
- Current progress – variants
- Next steps

Introduction to Khmer Language

- Khmer language has been written since the early 7th century using a script originating in South India
- Khmer borrowed some words from Sanskrit and Pāli
- Khmer was borrowed and found in Thai, Lao, Kuay, Stieng, Samre, Cham and others
- Official language in Cambodia with 15 million people
- 1.3 million people in southeastern Thailand
- More than a million people in southern Vietnam

Source: http://www.britannica.com/topic/Khmer-language

Introduction to Khmer Script

- Abugida Type
- Time period from c. 611–present
- System derived from Brahmi
- Thai and Lao derived from Khmer Script
- ISO 15924 - Khmr 355
  - Direction: Left-to-right
  - 146 Characters
- Unicode range
  - U+1780–U+17FF Khmer
  - U+19E0–U+19FF

Source - https://en.wikipedia.org/wiki/Khmer_alphabet

Current Membership of Khmer GP

Position	Name	Organization
Chair	Sopheap Seng	National Institute of Posts, Telecoms and ICT (NIPTICT), Cambodia
Secretary	Rapid Sun	Center of Research and Development, NIPTICT
Member	Daro Chin	Telecom Cambodia
Member	An Ra	Ministry of Post and Telecommunications
Member	Hong Danh	
Member	Ken Rangsey	Royal University of Phnom Penh
Member	Yatal Lim	Telecom Regulator of Cambodia
Member	Mok Khemera	Ministry of Posts and Telecommunications
Member	Than Makara	R & D Center, NIPTICT
Member	Chhan Kimsoeun	Royal University of Phnom Penh

Current Progress – Code Point Repertoire

- Consonants – Completed
- Independent Vowels – Completed
- Dependent Vowels – Completed
- Various Signs – Completed
- Lunar Date sign – Completed
- Currency Symbol – Completed
- Digits – Completed
- Numeric Symbol – Completed

Current Progress – Variants

- Define variant principle
  - 5 form/ position ex: ក្បឿង ប៊ិច
  - 4 styles
- Variant to Thai and Myanmar Language
- In progress – 80%

Next Steps

Activity	Description	Start Date	Status
Develop Principles	Principles to be used to determine valid code points, variants and labels	10 June 2015	100%
Determine Code Points	Select the code points from MSR which are needed for Root Zone LGR	10 July 2015	100%
Determine (any) Variants	From the codes points selected, determine if the end-user may confuse two code points	10 September 2015	80%
Determine Label Rules	Determine if there are any label level constrains on the use of selected code points	10 November 2015	--
Hold Public Consultation	Hold a workshop on the work accomplished by the generation panel to get feedback from the community and experts	Early December 2015	--
Write Proposal and Create XML	Write up the Root Zone LGR proposal, including references to each code point included, why variants needed and details of label rules developed + XML file	10 December 2015	--
Submit	Get public comments, finalize and submit	10 February 2016	--

Community Update:
Thai Generation Panel

Wanawit Ahkuputra
Chair, Thai GP

Internet in Thailand

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

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

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

Thai Script

1

ISO 15924

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

2

Unicode Range:

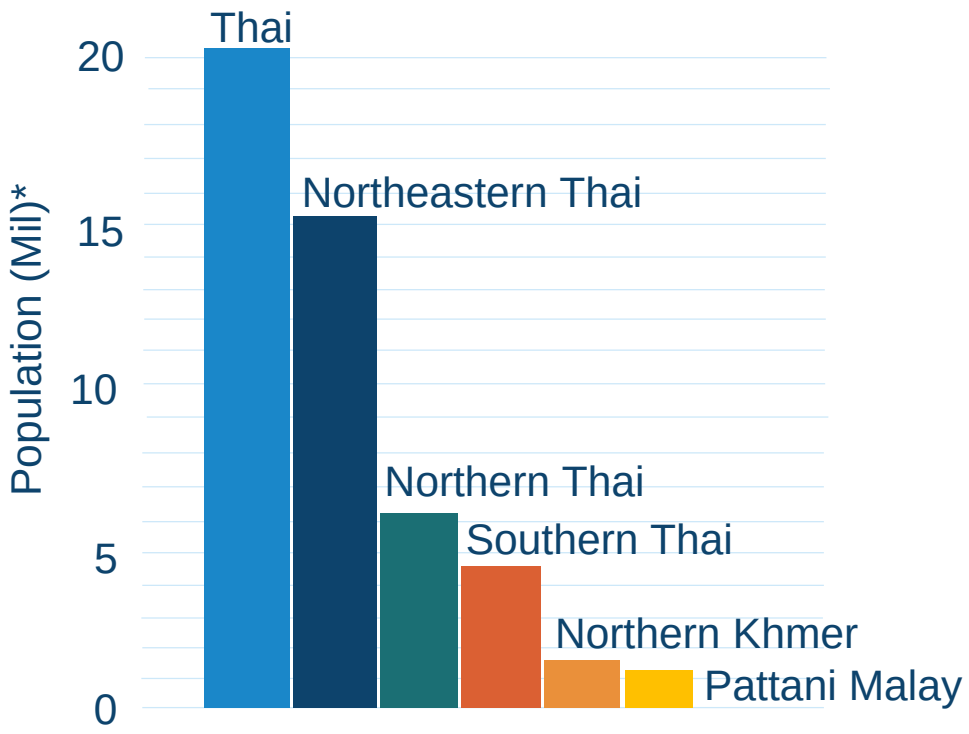
U+0E00 – U+0E7F

3

Writing systems that use Thai script

35 languages

Selected languages written in **Thai script**

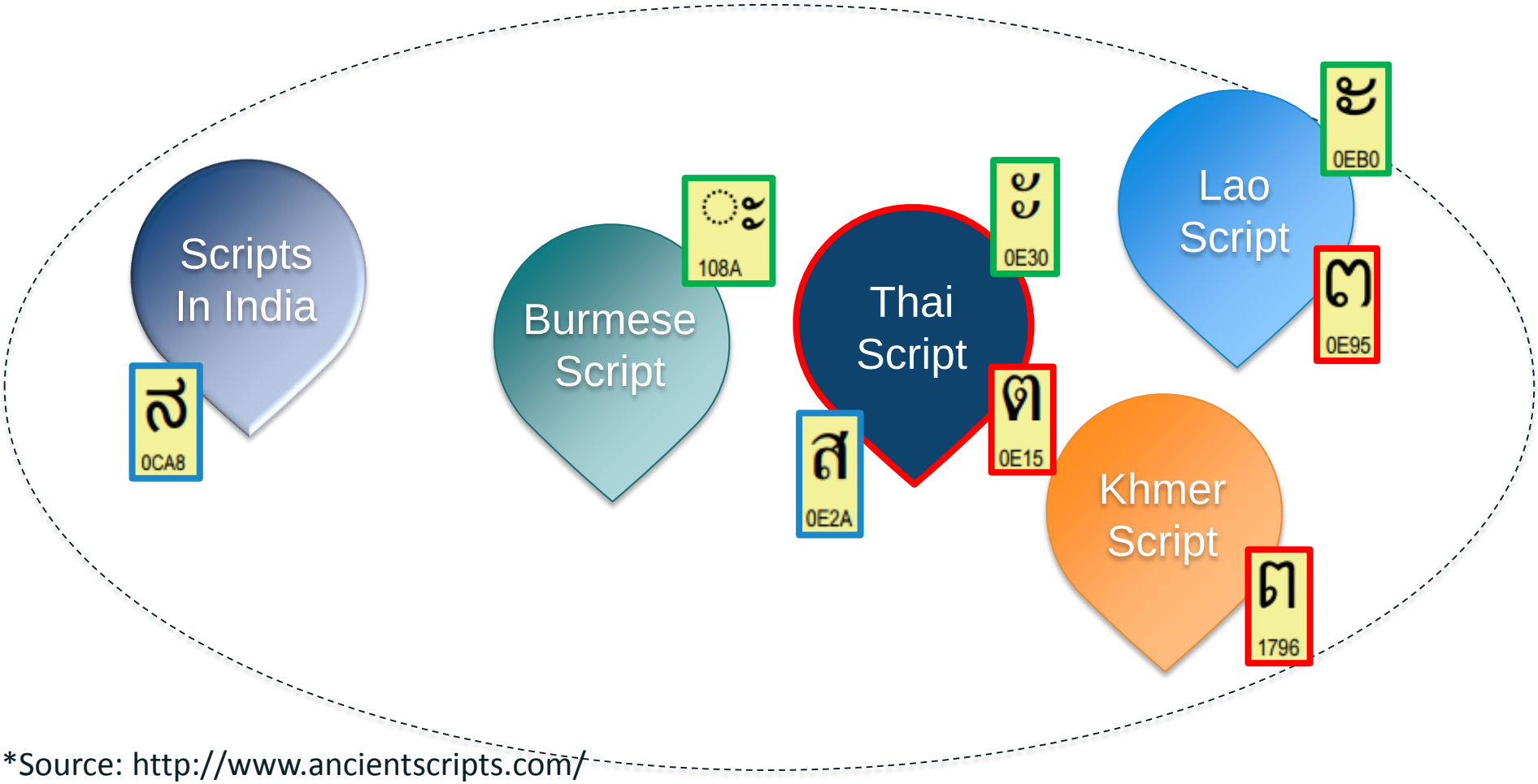


*Source: www.ethnologue.com and scriptsource.org

Thai Script and its Variants

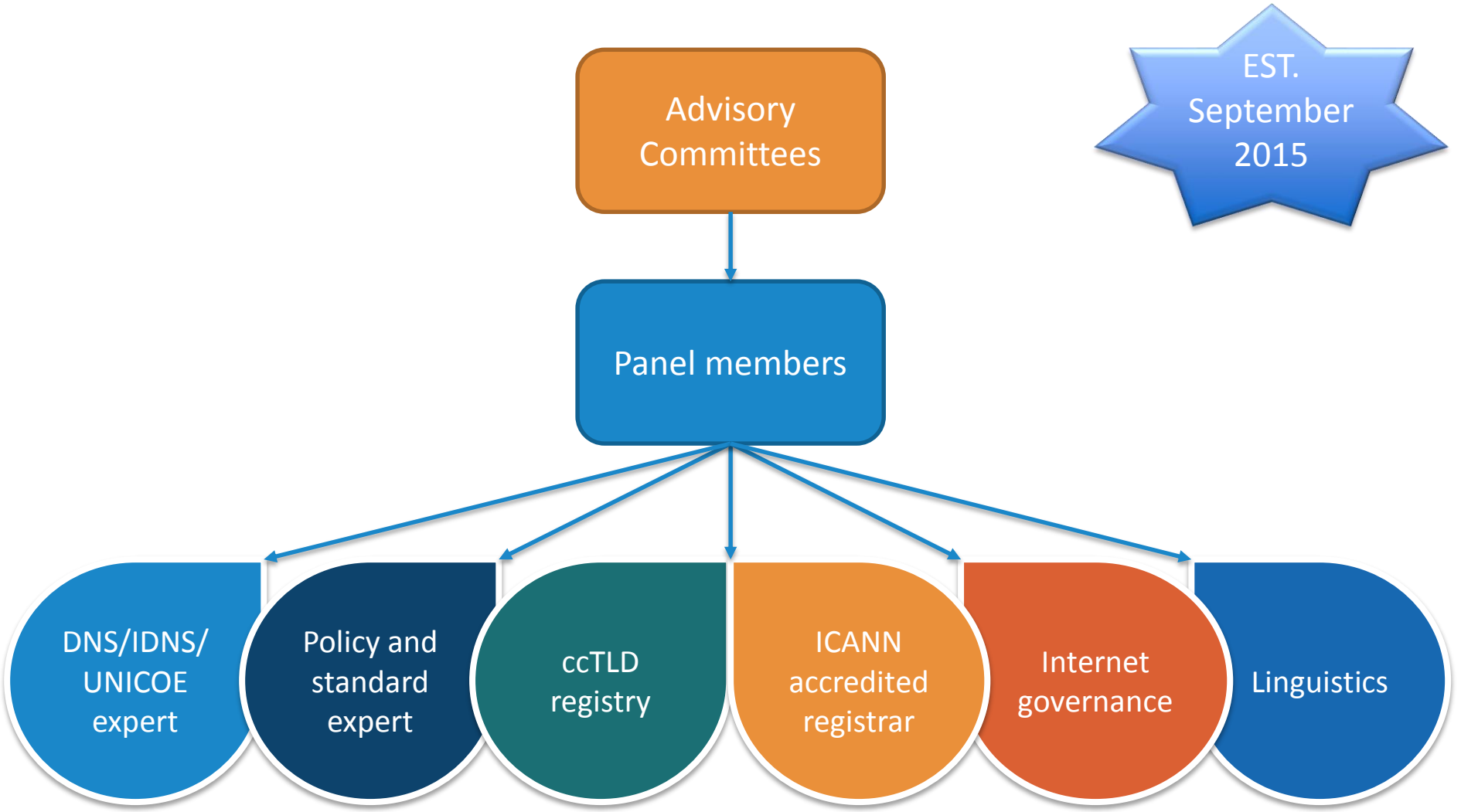
Brahmi script
Khmer script

Thai script*

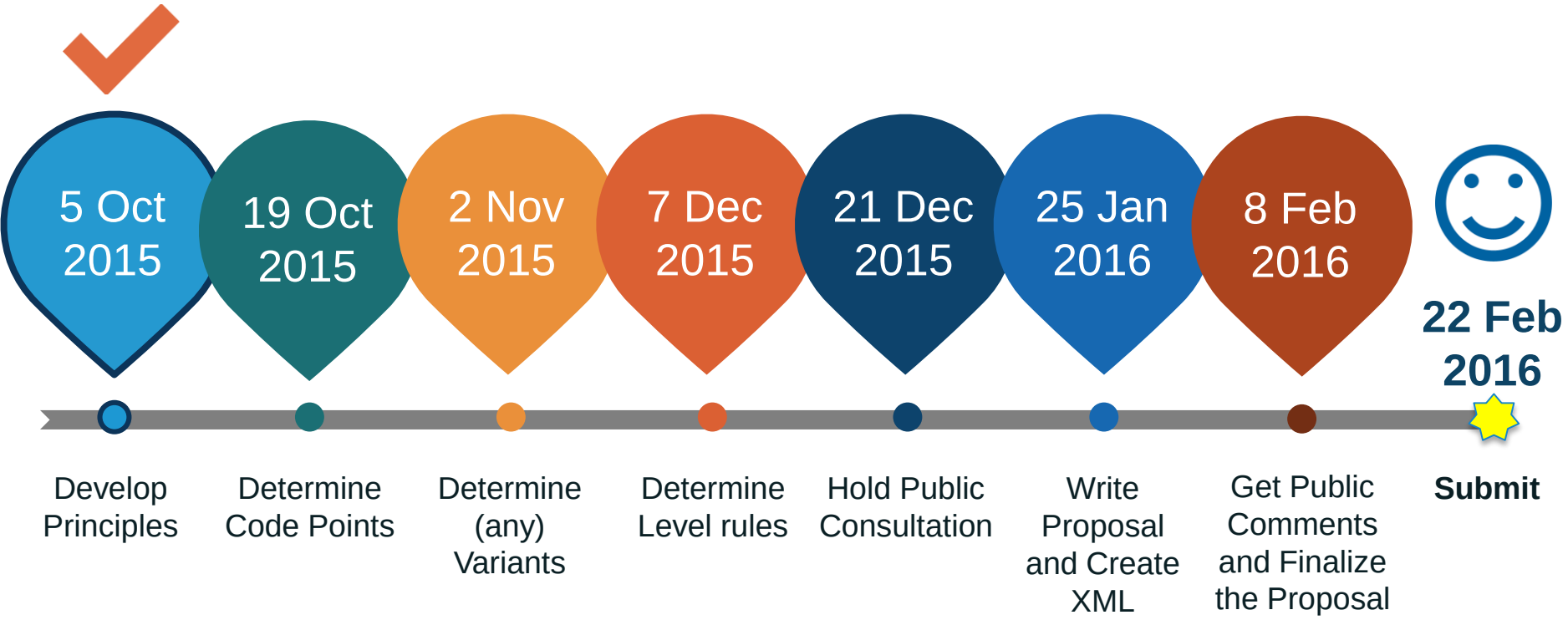


*Source: http://www.ancientscripts.com/

Thai Generation Panel



Timeline Thai Script LGR



To Summarize

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

Thank You

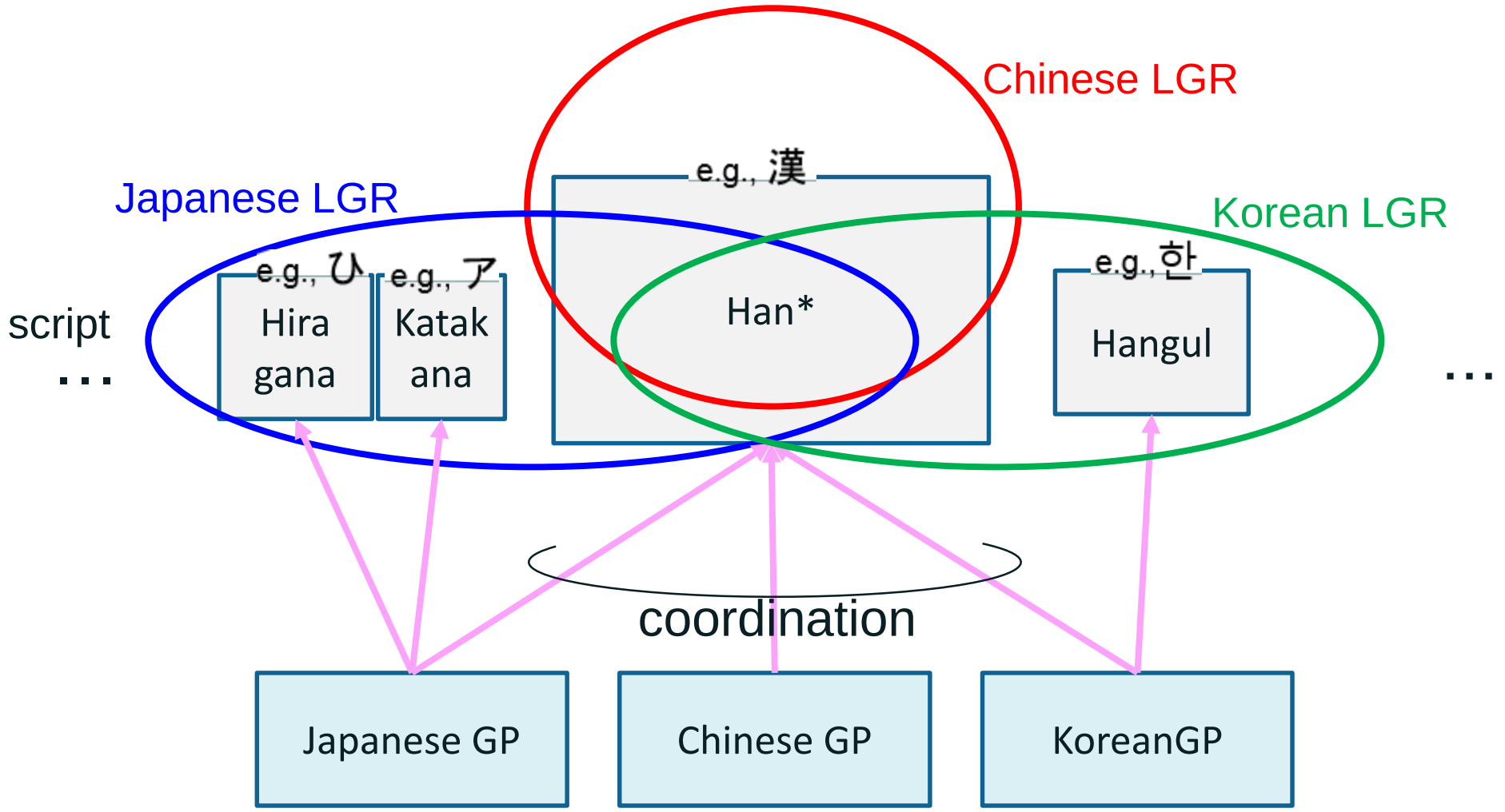
Wanawit Ahkuputra

wanawit@etda.or.th
Electronic Transactions Development Agency (Public Organization)
THAILAND

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

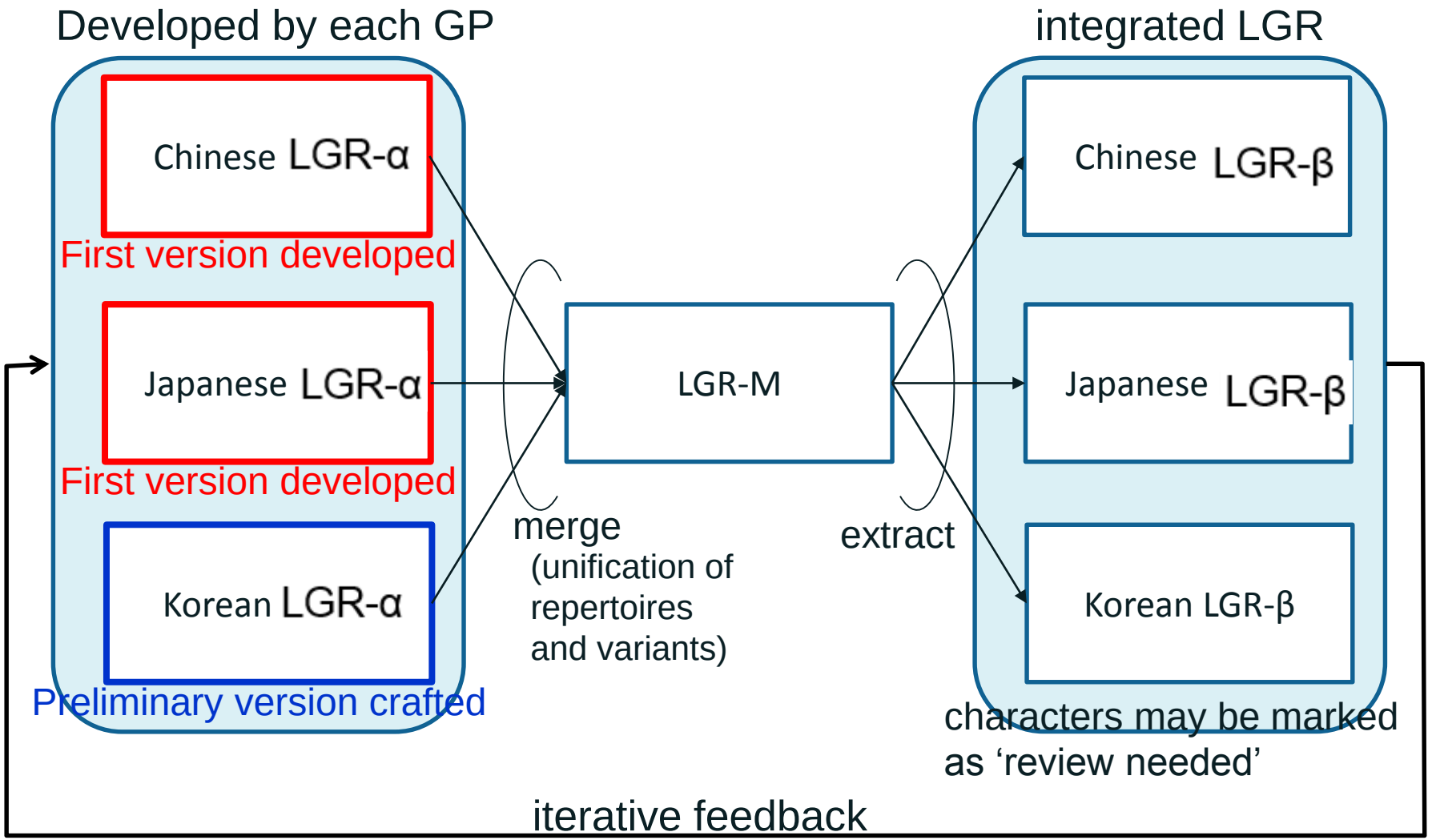
Typical Issues (especially re. Han Characters)

- Each of CJK has thousands of Han characters
  - MSR has about 20,000 Han characters
  - CGP picks up about 19,000 Han characters from MSR
  - JGP picks up about 6,000 Han characters from MSR
  - KGP picks up about 5,000 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 identifies 37 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 early 2015
  - CGP and JGP met during IETF in March 2015
- Coordination committees (formal)
  - CGP, JGP, and KGP meeting
    - for 1.5 days in May
    - four times during June ICANN Buenos Aires meeting
    - a couple of times during October ICANN Dublin meeting
  - More meetings needed to coordinate and conclude
  - Conclusion expected to 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
(revised by agreement in Buenos Aires)



Top-Ranked Discussion Items

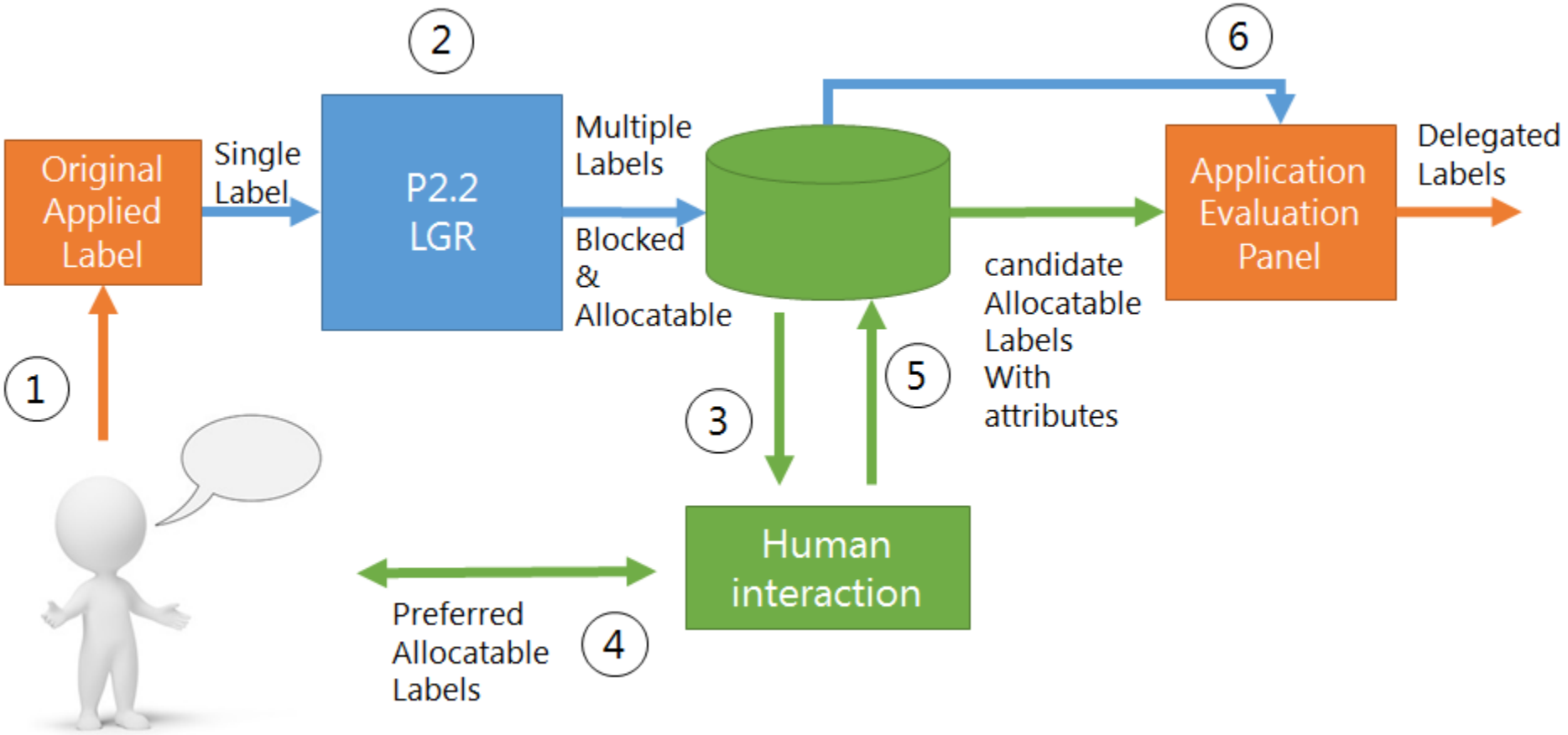
- Limiting the number of allocatable variant labels
  - Reduction of variant characters
  - Devising WLEs with a dedicated/an amended definition of variant subtypes and rules

    ➔ No easy solution found

    ➔ May need some artifice as proposed below
- Investigation of the possibility of using RootLGR as a process element to gain more flexibility
  - A proposal for coordination between the Root LGR and human intervention (i.e., application evaluation panel) is being discussed.

    ➔ Proposals will be pondered and sent to ICANN

Process Revision of Complementing Root LGR?



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