
ICANN
ANNUAL GENERAL
60
ABU DHABI
28 October–3 November 2017

Latin Generation Panel Meeting

Latin GP

ICANN 60
31 October 2017

Overview of Session Presentations

- Latin Generation Panel Overview - Mirjana Tasić
- Latin Repertoire Group - Mats Dufberg
- Latin Variant Working Group - Bill Jouris
- Q/A

Latin Generation Panel Overview

Mirjana Tasić
Latin GP Chair

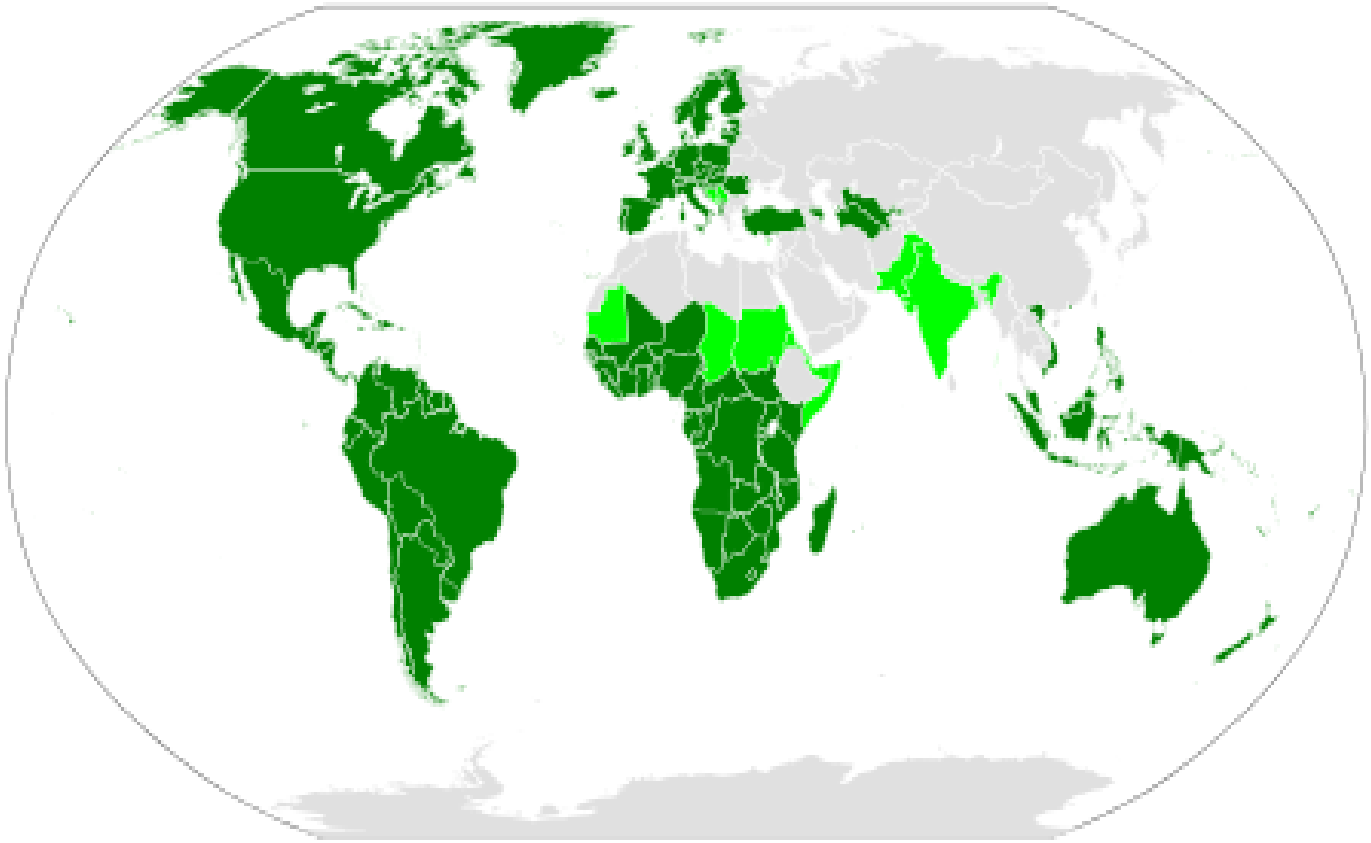
Latin GP Overview - Introduction

1. Short History
2. Scope of Work
3. Geographic and Linguistic Spread
4. Members
5. Challenges with Scope and Solutions
6. Challenges with Membership and Solutions

Latin GP Overview - Short History

- Summer 2016 - GP restarted with new call for volunteers
- Real work started October 2016
- GP proposal finalized and sent to ICANN at the beginning of May 2017
- GP seated on Monday, 15 May 2017
  - Proposal for Formation of Latin Generation Panel

Latin Script Geographic and Linguistic Spread



The dark green areas show the countries where the Latin script is the sole main script.
Light green shows countries where Latin co-exists with other scripts.
Grey areas - Latin-script alphabets are sometimes extensively used in areas colored grey due to the use of unofficial second languages, such as French in Algeria and English in Egypt, and to Latin transliteration of the official script, such as pinyin in China or ***rōmaji*** in Japan.

Latin GP Overview – Scope of Work

- Maximal Starting Repertoire version 2 (MSR-2)
- Lowercase letters
- Unicode ranges
  - Controls and Basic Latin
  - Controls and Latin-1 Supplement
  - Latin Extended-A
  - Latin Extended-B
  - IPA Extensions
  - Combining Diacritical Marks
  - Combining Diacritical Marks Supplement
  - Latin Extended Additional
  - Latin Extended-C
- Non exhaustive list of 455 languages in scope
- Non exhaustive list of EGIDS 1-5 languages contains 300 languages
- Non exhaustive list of EGIDS 1-4 languages contains 180 languages
- Maximal Starting repertoire version 2 (MSR-2) shows 279 Latin script code points

Note: **EGIDS** stands for the Expanded Graded Intergenerational Disruption Scale. This is a tool that is used to measure the status of a language in terms of endangerment or development.

Latin GP Overview – Members

- 14 members, 3 observers
- Language representatives
  - Africa
  - Asia
  - Australia and Oceania
  - Europe
  - North America
- Diversity
  - Community Representatives
  - Linguistic Experts
  - Registry/Registrar Experts
  - Policy Experts
  - Technical Community, DNS Experts
  - IDNA/Unicode Experts

Challenges with Scope and Solutions

- Challenges
  - Many languages
  - Many code points to process
- Solutions
  - Process languages with EGIDS=1-4 first (180)
  - Consider processing languages with EGIDS=5 (120)
  - Define simple procedure for developing Latin Script Repertoire

Challenges with Membership and Solutions

- Challenges
  - Representatives from different geographic regions
  - Not enough members to cover workload
- Solutions
  - Contacting people during different ICANN venues
  - Workload divided in two groups
    - Repertoire Group
    - Variant Group

Latin GP Overview - Current status

1. Overview of The Plan
2. Organization of Working Groups
3. Work Accomplished
4. Unofficial Public Comment on Principles
5. Tentative Completion Date

Latin GP Overview - Work Plan

Step	Activity	Duration
1	Initiate work	2 weeks
2	Develop principles	4 weeks
3	Develop Repertoire and Variants in parallel	24 weeks
4	Integrate results from two working groups	4 weeks

Latin GP Overview - Work Plan

Step	Activity	Duration
5	Discuss WLE rules needed for Latin script LGR	6 weeks
6	Prepare Latin script LGR proposal for public comment	8 weeks
7	Finalize Latin LGR proposal	10 weeks
Total Time: From Start to End		58 weeks

Latin GP Overview – Organization of Working Groups

- Repertoire Working group
  - 10 members
  - Developing Principles for Inclusion and Exclusion of Code Points in Latin Script for the Root Zone LGR
  - Processing Languages to build the repertoire
- Variant Working Group
  - 4 members
  - Developing Principles for Analysis of Variants in the Latin Script for the Root Zone LGR
  - Identifying variants

Latin GP Overview – Work Accomplished

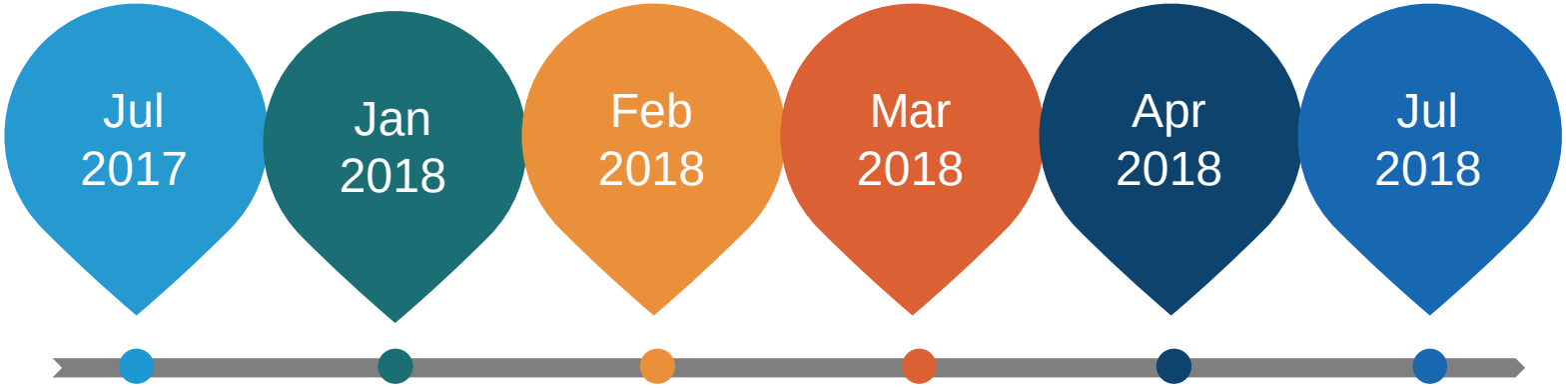
- Developing Repertoire
  - 180 of 180 EGIDS 1- 4 languages processed
  - 114 of 279 MSR-2 code points attested
  - 38 non-MSR-2 code points or code point sequences detected
- Developing Variants
  - Variants with Cyrillic script - work in progress
  - Variants with Greek script - to be done
  - Variants with any other script - to be done
  - In-script variants - to be done

Note: Figures as of 24 October 2017

Latin GP Overview – Unofficial Public Comment

- Unofficial consultations started on 13 September 2017
- Asking communities for comments on proposed principles
  - Principles for inclusion and exclusion of code points in Latin Script for the Root Zone LGR
  - Principles for Analysis of Variants in the Latin Script for the Root Zone LGR
- Expectations
  - Improvement of Principles

Latin GP Overview – Project Timeline



Jul 2017	Jan 2018	Feb 2018	Mar 2018	Apr 2018	Jul 2018
Developing Principles	Developing Code points Repertoire and Identifying Variants	Integration of different groups work products	Discuss WLE Rules needed for Latin Script LGR	Prepare Latin Script LGR Proposal for Public Comment	Submission to ICANN

To Summarize: work started on 15 May 2017
14 months is estimated time for project duration

Latin GP - Repertoire Working Group (WG) Overview

Mats Dufberg

Latin GP – Repertoire WG Overview

1. Repertoire Group Introduction
2. Scope of Work
3. Principles for Inclusion of Code Points
4. Principles for exclusion of code points
5. Organization of work
6. Challenges in creating Repertoire

Latin GP – Repertoire WG Introduction

- Repertoire Working Group (WG) is organized according to the Latin GP work plan
- 10 active participants
- Deadline for completing Repertoire – end of 2017

Latin GP – Repertoire WG Scope of Work

- Defining Principles for Inclusion and Exclusion of Latin Code Points in the Repertoire
- Processing of 180 EGIDS = 1-4 languages using Latin script in the first round to produce Repertoire
- Processing of language implies that every code point in specific language should be inspected and verified against MSR-2
- Status of processed code point
  - Code point is in MSR-2
  - Code point is not in MSR-2 -- noted for further discussion
  - Code point sequences -- noted for further processing

Note: Repertoire consists of all Latin script letters found in processed languages

Latin GP – Repertoire WG Principles

- Principles for Inclusion of code points in the Repertoire
  - Languages which have a rating of levels of 0-4 under the Expanded Graded Intergenerational Disruption Scale (EGIDS) are automatically considered as supporting the inclusion of a Code Point
  - Languages with EGIDS 5 may be included in special cases, where there is additional evidence that the language is in widespread use, notwithstanding its formal EGIDS rating
  - Code Points may only be included if they have established contemporary use in one or more of the languages considered

Latin GP – Repertoire WG Principles (cont.)

- Principles for Inclusion of code points in the Repertoire
  - If the Code Point in question is a Mark Code Point, then it can only be included in its context. That is, a Mark Code Point is included as part of a sequence consisting of a Lower Letter (Ll) or Other Letter (Lo) and the subsequent mark or marks
  - Any combination of Code Points is defined by its sequence. To be included, a sequence must be supported by some included language in the same way as a separate Code Point of type Ll or Lo
  - If a character can be represented by multiple Code Point Sequences, each Code Point Sequence must be separately justified to be included

Latin GP – Repertoire WG Principles (cont.)

- Principles for Inclusion of code points in the Repertoire
  - A Code Point Sequence can only be included if there is no pre-composed alternative available, unless there is specific evidence that a language eligible for inclusion under Criteria 1 makes alternate use of such a sequence
  - If the Code Point in question is a Modifier letter (Lm), then it can only be included together with its context. That is, a sequence of Lm plus Ll or Lo (or the other way around), unless there is strong evidence that the Lm can be used in any context, or that such a sequence or order cannot be defined

Latin GP – Repertoire WG Principles (cont.)

- Principles for Exclusion of code points from the Repertoire

  A Code Point is excluded if at least one of these exclusion principles is met. If a Code Point can neither be included nor excluded on the basis of these principles, the Code Point is automatically excluded from the proposed LGR for Latin Script, per RFC 6912

  - The Code Point is DISALLOWED or UNASSIGNED by IDNA 2008 protocol
  - The Code Point presents a security or stability issue which cannot be resolved at any other stage of the analysis (e.g., stage of determining Code Points, variants, Contextual Rules or Whole Label Evaluation Rules)
  - The Code Point is either deprecated or not recommended for use in Unicode Standard --unless it meets all of the applicable inclusion criteria, with no alternative Code Point or Code Point sequence

Latin GP – Repertoire WG Principles (cont.)

- Principles for Exclusion of code points from the Repertoire
  - The Code Point is used exclusively in a subset of textual genres, such as technical or religious texts, and is not otherwise used as described in Section 2 above
  - The Code Point is predominantly used in one of the following functions, apart from any other uses in orthography
    - Formatting character or mark
    - Numerical digit
    - Punctuation mark
    - Honorific mark or symbol
    - Mathematical symbol

Latin GP – Repertoire WG Organization of Work

- Resources for work
  - Language table with information about 180 languages EGIDS=1-4
  - Unicode table, https://unicode-table.com/en/#basic-latin
  - MSR-2 table with MSR-2 code points
  - Instructions on how to process languages
- Steps in processing specific language
  - Repertoire group member (RGM) picks one language from language table and notes in the Language table that specific language is in processing status
  - Language name column in Language table contains a link to omniglot.com entry for that language
  - RGM inspects every code point in the alphabet found on omniglot.com for that language

Latin GP – Repertoire WG Organization of Work

- Steps in processing specific language – continued from previous slide
  - If RGM finds it necessary, some other sources about the alphabet of the language in processing could be consulted
  - For finding UNICODE code points for the specific letter following link is used http://unicode.org/cldr/utility/character.jsp
  - RGM writes verification information for the specific code point in MSR-2 table
  - RGM writes her/his name in the Language table after processing of language is finished
  - RGM puts his/her initials in MSR-2 table row with verified code point

Latin GP – Repertoire WG Organization of Work

- Language table sample

	Language	ISO 639-3	Classification	Population	EGIDS	Language map	Processed by
1.	Afrikaans.	afr	Indo-European. Germanic. West. Low Saxon-Low Franconian. Low Franconian	7,096,810	1	Botswana. Lesotho. South Africa and Swaziland Namibia	Finished (Fiammetta Caccavale)
2.	Albanian. Arbëreshë Albanian [aae] (Italy) Arvanitika Albanian[aat] (Greece) Gheg Albanian [aln] (Serbia) Tosk Albanian [als]	sqi	Indo-European. Albanian.	5,367,000	1	Albania. Greece. Serbia. Macedonia. Montenegro	Finished (Mats Dufberg)
3.	Azeri.Azerbaijani	azj	Turkic. Southern. Azerbaijani	24,226,940	1	Azerbaijan.Georgia.Iraq Jordan and Syria	Finished (Mirjana Tasic)
4.	Chamorro.Chamorru Tjamoro	cha	Austronesian Malayo-Polynesian Chamorro	94,700	1	Guam and Northern Mariana Islands	Finished (Bakiau)
5.	Croatian.Hrvatski	hrv	Indo-European Balto-Slavic Slavic South Western	5,609,290	1	Croatia	Finished (Sarmad Hussain)
6.	Czech Bohemian Cestina	ces	Indo-European Balto-Slavic Slavic West Czech-Slovak	10,619,340	1	Czech Republic	Finished (Bill Jouris)

Latin GP – Repertoire WG Organization of Work

- MSR-2 table sample

Data exception	In MSR	Unicode	Glyph	Unicode name	Languages using the code point	Reference supporting inclusion	Initials
	X	0294	ʔ	LATIN LETTER GLOTTAL STOP			
	X	026A	ɪ	LATIN LETTER SMALL CAPITAL I			
	X	0061	a	LATIN SMALL LETTER A	English (1) Malay (1) Vietnamese (1) Kirundi (1) Turkish (1)	http://www.omniglot.com/writing/english.htm http://www.omniglot.com/writing/malay.htm http://www.omniglot.com/writing/vietnamese.htm http://www.omniglot.com/writing/kirundi.php http://www.omniglot.com/writing/turkish.htm	
	X	00E1	á	LATIN SMALL LETTER A WITH ACUTE	Spanish (1)	https://www.icann.org/sites/default/files/packages/lgr/lgr-second-level-spanish-30aug16-en.html	MB
	X	00E1	á	LATIN SMALL LETTER A WITH ACUTE	Czech (1)	http://www.omniglot.com/writing/czech.htm	BJ
	X	00E1	á	LATIN SMALL LETTER A WITH ACUTE	Icelandic (1)	http://www.omniglot.com/writing/icelandic.htm	BJ
	X	00E1	á	LATIN SMALL LETTER A WITH ACUTE	Faroese (2)	http://www.omniglot.com/writing/faroese.htm	BJ
	X	00E1	á	LATIN SMALL LETTER A WITH ACUTE	Kirundi (1)	https://en.wikipedia.org/wiki/Burundi_Bwacu#Kirundi_.28with_tonal_diacritics_.E2.80.94_utw.C3.A2tuzo.29	JPN
	X	00E1	á	LATIN SMALL LETTER A WITH ACUTE	Chuukese (2)	http://www.omniglot.com/writing/chuukese.htm	MM
	x	00E1	á	LATIN SMALL LETTER A WITH ACUTE	Galician (2)	http://www.webcitation.org/6siTI8ieQ	MM
	x	00E1	á	LATIN SMALL LETTER A WITH	Lule Sámi (2)	http://www.omniglot.com/writing/lulesami.htm	MT

Challenges in Developing Repertoire

- Should we process some languages with EGIDS = 5 ?
  - Selecton criteria
    - No of speakers ?
    - Some other criteria
    - ???

SOME MORE TOPICS TO DISCUSS

- How to treat code point sequences?
  - 0075 + 0322 (LATIN SMALL LETTER U + COMBINING RETROFLEX HOOK BELOW) in Lithuanian
  - 2019 + 0077 (RIGHT SINGLE QUOTATION MARK + LATIN SMALL LETTER W) in Dagaare - Burkina Faso

Challenges in Developing Repertoire (cont.)

SOME MORE TOPICS TO DISCUSS

- Should we consider consonant digraphs or trigraphs?
  - ny, nyh, dh - like in Wa (Va) language
- Click consonants -- how to threat them?
- What to do with letters not specified in MSR and found in processed languages
  - 0264 (ɤ) LATIN SMALL LETTER RAMS HORN in Aklan (4)
- Treatment of 0027, APOSTROPHE
- Treatment of MODIFIER LETTERS
  - 02BB MODIFIER LETTER APOSTROPHE
  - 02BC, MODIFIER LETTER TURNED COMMA
  - ……

Challenges in Developing Repertoire (cont.)

SOME MORE TOPICS TO DISCUSS

- What to do when a symbol is used for a language, but does not appear to have a Unicode code point at all? Nor any way to create one, using the available Unicode symbols.

  Example: Tuvan language
  a в c d e f g ɑ̨ i j k l m n □ o ө p r s ş t u v x y z ƶ ь

- …….

Latin GP – Variant Working Group (WG)

Bill Jouris

Latin GP Variant WG - Analysis of Variants

- Objective
  - Determine variant relationships with related scripts, such as Cyrillic, Greek and Armenian, and others as the panel sees fit
  - Analyze variant relationships within Latin script
- Scope
  - MSR-2
  - Latin GP Repertoire
- Principles
  - Relationship of two letters (or combination of letters) is sufficiently universal across the entire Latin script community
  - Code points (or sequences) are visually identical

Latin GP Variant WG - Analysis of Variants

- Examples for analysis:

Visually Identical	These code points are visually identical: Latin Small Letter E ‘e’ (0065) Cyrillic Small Letter IE ‘e’ (0435) These code points are not visually identical: Latin Small Letter E ‘e’ (0065) Latin Small Letter E with diaeresis ‘ë‘ (00EB)
Orthographic Considerations	Eszett in German: Latin Small Letter S ‘ss’ (0073 0073) Latin Small Letter Sharp S ‘ß’ (00DF)
Normalization Exceptions	When two visually identical letters, albeit with different code point sequences do not normalize to the same canonical form

Latin GP Variant WG - Analysis of Variants

- **Cross-Script Analysis: Cyrillic**
  - Preliminary findings
    - Cyrillic proposal: 17 variant sets
    - Found additional identical glyphs in both repertoires (e.g., 0103 and 04D1 “ă ” , 0115 and 04D7 “ĕ”)

Latin GP Variant WG - Analysis of Variants

- **Next Steps**
  - Continue cross-script analysis
    - Greek
    - Armenian
  - Analyze variant relationships within Latin script as repertoire is developed

Questions / Issues

Variants Working Group

Latin Generation Panel

A Moving Target

- Early summer:

Cyrillic	Latin
04AB	00E7
	

- Mid-October:

Cyrillic	Latin
04AB	00E7
	

Upper and Lower Case

Is one of these different?

.com

.COM

.com

Fonts

Is one of these different?

.ьay

.bay

.bay

Variants vs. Mere “Confusables”

The underlying issue is this:

What is the real-world problem that we are trying to address here?

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