
ICANN
ANNUAL GENERAL
63
BARCELONA
20–25 October 2018

Latin Script Root Zone Label Generation Rules

Mats Dufberg
Michael Bauland
Mirjana Tasić
Latin GP

Agenda Overview

1. Scope of Work
2. Challenge and Solutions
3. Work Accomplished
4. Project Timeline
5. Variant Analysis Discussion
6. WLE and Test File Discussion

Scope of Work for Code Point Analysis

- Maximal String Repertoire Version 3 (MSR-3)
  - MSR-3 is a subset what is accepted for IDNA 2008
- Unicode ranges
  - Controls and Basic Latin
  - Controls and Latin-1 Supplement
  - Latin Extended-A only lowercase
  - 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 181 languages

Scope of Work Variant Analysis

- In-script variant analysis
- Cross-script variant analysis
  - Armenian script
  - Cyrillic script
  - Greek script

Challenges

- Challenges
  - Many languages
  - Many code points to process
  - Not enough members from too few regions to cover workload
  - Since Latin script is used by languages from different language families and geographic areas all over the world, a rich variety of characters have been developed to meet the need of representing different linguistic characteristics
  - The same abstract character in a specific languages can sometimes have different shapes that overlap with different code points, e.g. both {f} and {*ƒ*} can be used for letter “f” in Swedish, but they are also encoded as different code points in Unicode

Solutions

- Solutions
  - First process languages with EGIDS=1-4 (181)
  - Consider processing languages with EGIDS=5 (119)
    - Decided to limit level 5 languages to those with at least 1 million users and with sufficient reference
    - 29 level 5 languages are included in the investigation
  - Define simple procedure for developing Latin script repertoire
    - Go through available online resources for each languages looking for the “alphabet” for the language to determine what code points are need to support the languages.
  - Workload divided in two groups
    - Repertoire Working Group
    - Variant Working Group

Work Accomplished – Repertoire

- Developing Repertoire
  - 181 of 181 EGIDS 1- 4 languages processed
  - 29 EGIDS 5 languages processed
  - 195 of 279 MSR-2 code points attested
    - 192 letter code points, e.g. “ə” U+0259 LATIN SMALL LETTER SCHWA
    - 7 combining marks, e.g. U+0331 COMBINING MACRON BELOW
    - Many of the letter code points are pre-composed characters with diacritics, e.g. "ẍ" U+1E8D LATIN SMALL LETTER X WITH DIAERESIS

Work Accomplished – Repertoire

- Developing Repertoire
  - 3 non-MSR-2 code points were proposed and accepted in MSR-3

ɨ	0268	LATIN SMALL LETTER I WITH STROKE
ɲ	0272	LATIN SMALL LETTER N WITH LEFT HOOK
ḽ	1E3D	LATIN SMALL LETTER L WITH CIRCUMFLEX BELOW

Work Accomplished – Repertoire

- Developing Repertoire
  - 3 non-MSR 3 code points are proposed for inclusion in an updated MSR, and if accepted, they will be included in Latin LGR proposal

d̂	1E13	LATIN SMALL LETTER D WITH CIRCUMFLEX BELOW
n̂	1E4B	LATIN SMALL LETTER N WITH CIRCUMFLEX BELOW
t̂	1E71	LATIN SMALL LETTER T WITH CIRCUMFLEX BELOW

Work Accomplished – Repertoire

- Outside Repertoire
  - Some languages use code points that cannot be accepted for the root zone, e.g. apostrophe like characters

’	02BC	LATIN MODIFIER LETTER APOSTROPHE

  - Latin GP rejected a code point that is very similar to exclamation mark, but used in one of the investigated languages

!	01C3	LATIN LETTER RETROFLEX CLICK

Work Accomplished – Repertoire

- Outside Repertoire
  - Latin GP proposed 3 code points used for click sounds for inclusion in MSR 3, but IP did not accept them on the grounds that they are visually too similar to code points disallowed in IDNA 2008

ǀ	01C0	LATIN LETTER DENTAL CLICK
ǁ	01C1	LATIN LETTER LATERAL CLICK
ǂ	01C2	LATIN LETTER ALVEOLAR CLICK

Work Accomplished – Variants

- Developing Variants
  - Framework defined
    - In-script variants are under analysis
    - Cross-script variants with Cyrillic, Greek and Armenian: Preliminary list complete

Work Accomplished – IP feedback

- Latin GP submitted the second round preliminary proposal to the IP in September 2018
- The IP has responded with feedback, especially on cross-script variants
- Latin GP analysis of the feedback on the proposal will result in a new, complete proposal, that will eventually go for public comment

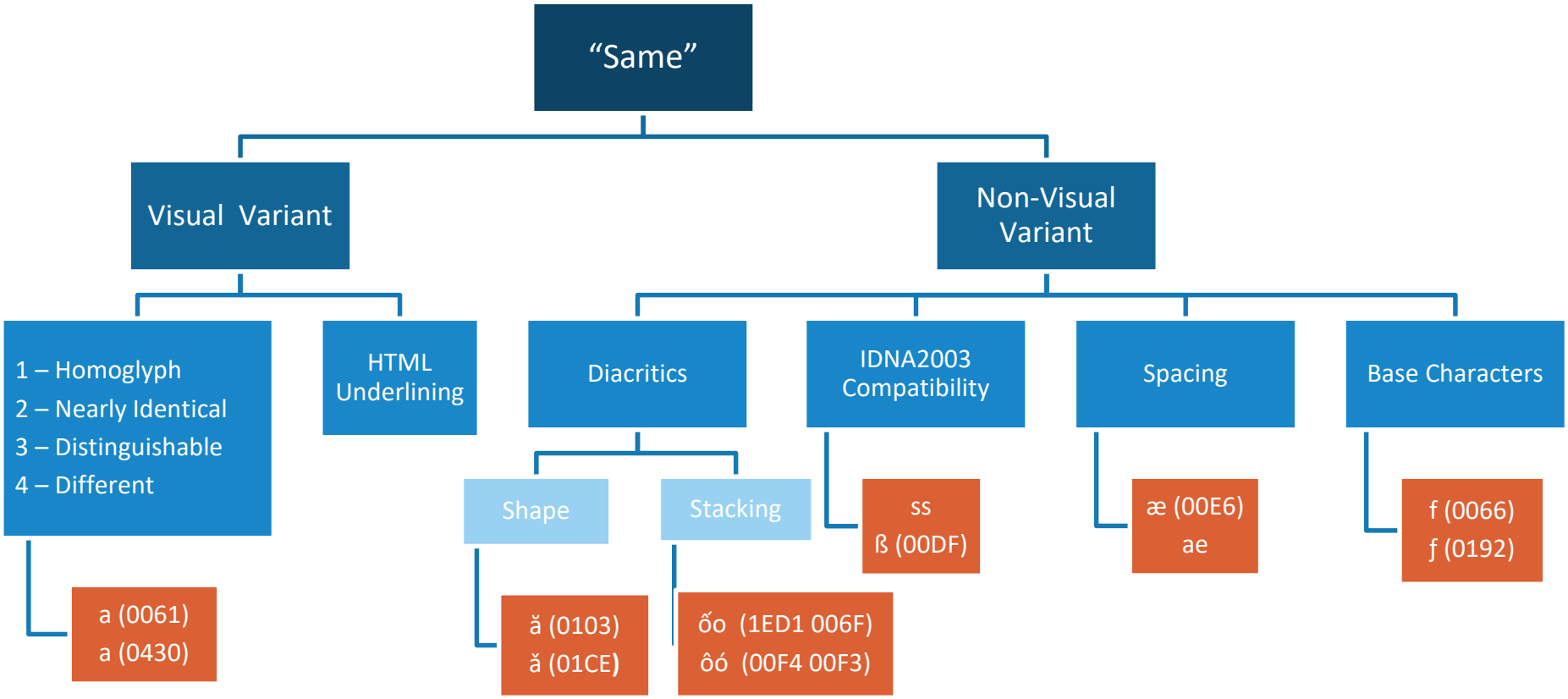
Project Timeline

Jul 2017	Jan 2018	May 2018	Sep 2018	Nov 2018	Dec 2018
Developing Principles	Developing code points repertoire and identifying variants	Review the code points repertoire with the IP	Review the variant rules with the IP	Finalize the Latin script LGR Proposal	Prepare Latin script LGR proposal for Public Comment

Variant Analysis

Cross-Script and In-Script

Tentative Variant Analysis Framework



Cross-Script Variants

Cyrillic, Greek and Armenian

Preliminary Candidates: Latin-Cyrillic (1/3)

Source Unicode Name	Source Code Point	Source Glyph	Target Glyph	Target Code Point	Target Unicode Name	Rationale	Proposed Cross-Script Variant by Cyrillic GP	Cross-Script Variant Candidate
LATIN SMALL LETTER A	0061	a	а	0430	CYRILLIC SMALL LETTER A	Homoglyph	Yes	Candidate
LATIN SMALL LETTER C	0063	c	с	0441	CYRILLIC SMALL LETTER ES	Homoglyph	Yes	Candidate
LATIN SMALL LETTER E	0065	e	е	0435	CYRILLIC SMALL LETTER IE	Homoglyph	Yes	Candidate
LATIN SMALL LETTER H	0068	h	h	04BB	CYRILLIC SMALL LETTER SHHA	Homoglyph	Yes	Candidate
LATIN SMALL LETTER I	0069	i	і	0456	CYRILLIC SMALL LETTER BYELORUSSIAN-UKRAINIAN I	Homoglyph	Yes	Candidate
LATIN SMALL LETTER J	006A	j	ј	0458	CYRILLIC SMALL LETTER JE	Homoglyph	Yes	Candidate
LATIN SMALL LETTER L	006C	l	ӏ	04CF	CYRILLIC SMALL LETTER PALOCHKA	Homoglyph	Yes	Candidate
LATIN SMALL LETTER O	006F	o	о	043E	CYRILLIC SMALL LETTER O	Homoglyph	Yes	Candidate
LATIN SMALL LETTER P	0070	p	р	0440	CYRILLIC SMALL LETTER ER	Homoglyph	Yes	Candidate
LATIN SMALL LETTER R	0072	r	г	0433	CYRILLIC SMALL LETTER GHE	Glyphs nearly identical due to font design		Candidate
LATIN SMALL LETTER S	0073	s	ѕ	0455	CYRILLIC SMALL LETTER DZE	Homoglyph	Yes	Candidate
LATIN SMALL LETTER X	0078	x	х	0445	CYRILLIC SMALL LETTER HA	Homoglyph	Yes	Candidate

Preliminary Candidates: Latin-Cyrillic (2/3)

Source Unicode Name	Source Code Point	Source Glyph	Target Glyph	Target Code Point	Target Unicode Name	Rationale	Proposed Cross-Script Variant by Cyrillic GP	Cross-Script Variant Candidate
LATIN SMALL LETTER Y	0079	y	ɣ	04AF	CYRILLIC SMALL LETTER STRAIGHT U	Glyphs nearly identical due to font design		Candidate
LATIN SMALL LETTER Y	0079	y	у	0443	CYRILLIC SMALL LETTER U	Homoglyph	Yes	Candidate
LATIN SMALL LETTER A WITH DIAERESIS	00E4	ä	ӓ	04D3	CYRILLIC SMALL LETTER A WITH DIAERESIS	Homoglyph	Yes	Candidate
LATIN SMALL LETTER AE	00E6	æ	æ	04D5	CYRILLIC SMALL LIGATURE A IE	Homoglyph	Yes	Candidate
LATIN SMALL LETTER C WITH CEDILLA	00E7	ç	ç	04AB	CYRILLIC SMALL LETTER ES WITH DESCENDER	Glyphs nearly identical due to font design		Candidate
LATIN SMALL LETTER E WITH DIAERESIS	00EB	ë	ё	0451	CYRILLIC SMALL LETTER IO	Homoglyph	Yes	Candidate
LATIN SMALL LETTER I WITH DIAERESIS	00EF	ï	ї	0457	CYRILLIC SMALL LETTER YI	Homoglyph		Candidate
LATIN SMALL LETTER O WITH DIAERESIS	00F6	ö	ӧ	04E7	CYRILLIC SMALL LETTER O WITH DIAERESIS	Homoglyph		Candidate
LATIN SMALL LETTER Y WITH DIAERESIS	00FF	ÿ	ӱ	04F1	CYRILLIC SMALL LETTER U WITH DIAERESIS	Glyphs nearly identical due to font design		Candidate
LATIN SMALL LETTER A WITH BREVE	0103	ă	ӑ	04D1	CYRILLIC SMALL LETTER A WITH BREVE	Homoglyph		Candidate

Preliminary Candidates: Latin-Cyrillic (3/3)

Source Unicode Name	Source Code Point	Source Glyph	Target Glyph	Target Code Point	Target Unicode Name	Rationale	Proposed Cross-Script Variant by Cyrillic GP	Cross-Script Variant Candidate
LATIN SMALL LETTER E WITH BREVE	0115	ĕ	ĕ	04D7	CYRILLIC SMALL LETTER IE WITH BREVE	Homoglyph		Candidate
LATIN SMALL LETTER H WITH STROKE	0127	ħ	ħ	045B	CYRILLIC SMALL LETTER TSHE	Homoglyph	Yes	Candidate
LATIN SMALL LETTER R WITH ACUTE	0155	ŕ	ŕ	0453	CYRILLIC SMALL LETTER GJE	Glyphs nearly identical due to font design		Candidate
LATIN SMALL LETTER TURNED E	01DD	ǝ	ǝ	04D9	CYRILLIC SMALL LETTER SCHWA	Homoglyph	Yes	Candidate
LATIN SMALL LETTER R WITH STROKE	024D	ɍ	ɍ	0493	CYRILLIC SMALL LETTER GHE WITH STROKE	Glyphs nearly identical due to font design		Candidate
LATIN SMALL LETTER SCHWA	0259	ə	ə	04D9	CYRILLIC SMALL LETTER SCHWA	Homoglyph	Yes	Candidate
LATIN SMALL LETTER EZH	0292	ʒ	ʒ	04E1	CYRILLIC SMALL LETTER ABKHASIAN DZE	Homoglyph		Candidate
LATIN SMALL LETTER U WITH DOT BELOW	1EE5	ụ	ụ	045F	CYRILLIC SMALL LETTER DZHE	Glyphs nearly identical due to font design		Candidate
LATIN SMALL LETTER Y WITH TILDE	1EF9	ỹ	ȳ	04EF	CYRILLIC SMALL LETTER U WITH MACRON	Glyphs nearly identical due to font design		In Review

Preliminary Candidates: Latin-Greek (1/2)

Source Unicode Name	Source Code Point	Source Glyph	Target Glyph	Target Code Point	Target Unicode Name	Rationale	Cross-Script Variant Candidate
LATIN SMALL LETTER A	0061	a	α	03B1	GREEK SMALL LETTER ALPHA	Glyphs nearly identical due to font design	Candidate
LATIN SMALL LETTER O	006F	o	ο	03BF	GREEK SMALL LETTER OMICRON	Homoglyph	Candidate
LATIN SMALL LETTER P	0070	p	ρ	03C1	GREEK SMALL LETTER RHO	Glyphs nearly identical due to font design	Candidate
LATIN SMALL LETTER U	0075	u	υ	03C5	GREEK SMALL LETTER UPSILON	Glyphs nearly identical due to font design	Candidate
LATIN SMALL LETTER V	0076	v	ν	03BD	GREEK SMALL LETTER NU	Glyphs nearly identical due to font design; based on security	Candidate
LATIN SMALL LETTER X	0078	x	χ	03C7	GREEK SMALL LETTER CHI	Glyphs nearly identical due to font design	In Review
LATIN SMALL LETTER Y	0079	y	γ	03B3	GREEK SMALL LETTER GAMMA	Glyphs nearly identical due to font design	Candidate
LATIN SMALL LETTER SHARP S	00DF	ß	β	03B2	GREEK SMALL LETTER BETA	Glyphs nearly identical due to font design	Candidate
LATIN SMALL LETTER A WITH ACUTE	00E1	á	ά	03AC	GREEK SMALL LETTER ALPHA WITH TONOS	Glyphs nearly identical due to font design	Candidate
LATIN SMALL LETTER I WITH ACUTE	00ED	í	ί	03AF	GREEK SMALL LETTER IOTA WITH TONOS	Homoglyph	Candidate
LATIN SMALL LETTER I WITH DIAERESIS	00EF	ï	ϊ	03CA	GREEK SMALL LETTER IOTA WITH DIALYTIKA	Homoglyph	Candidate
LATIN SMALL LETTER O WITH ACUTE	00F3	ó	ό	03CC	GREEK SMALL LETTER OMICRON WITH TONOS	Homoglyph	Candidate

Preliminary Candidates: Latin-Greek (2/2)

Source Unicode Name	Source Code Point	Source Glyph	Target Glyph	Target Code Point	Target Unicode Name	Rationale	Cross-Script Variant Candidate
LATIN SMALL LETTER U WITH ACUTE	00FA	ú	ύ	03CD	GREEK SMALL LETTER UPSILON WITH TONOS	Glyphs nearly identical due to font design	Candidate
LATIN SMALL LETTER U WITH DIAERESIS	00FC	ü	ϋ	03CB	GREEK SMALL LETTER UPSILON WITH DIALYTIKA	Glyphs nearly identical due to font design	Candidate
LATIN SMALL LETTER DOTLESS I	0131	ı	ι	03B9	GREEK SMALL LETTER IOTA	Homoglyph	Candidate
LATIN SMALL LETTER O WITH HORN	01A1	ơ	σ	03C3	GREEK SMALL LETTER SIGMA	Glyphs nearly identical due to font design	Candidate
LATIN SMALL LETTER OPEN E	025B	ɛ	ε	03B5	GREEK SMALL LETTER EPSILON	Homoglyph	Candidate
LATIN SMALL LETTER IOTA	0269	ɩ	ι	03B9	GREEK SMALL LETTER IOTA	Homoglyph	Candidate
LATIN SMALL LETTER V WITH HOOK	028B	ʋ	υ	03C5	GREEK SMALL LETTER UPSILON	Glyphs nearly identical due to font design	Candidate

Preliminary Candidates: Latin-Armenian (1/1)

Source Unicode Name	Source Code Point	Source Glyph	Target Glyph	Target Code Point	Target Unicode Name	Rationale	Cross-Script Variant Candidate
LATIN SMALL LETTER G	0067	g	g	0581	ARMENIAN SMALL LETTER CO	Glyphs nearly identical due to font design	Candidate
LATIN SMALL LETTER H	0068	h	h	0570	ARMENIAN SMALL LETTER HO	Glyphs nearly identical due to font design	Candidate
LATIN SMALL LETTER N	006E	n	n	0578	ARMENIAN SMALL LETTER VO	Glyphs nearly identical due to font design	Candidate
LATIN SMALL LETTER O	006F	o	o	0585	ARMENIAN SMALL LETTER OH	Homoglyph	Candidate
LATIN SMALL LETTER Q	0071	q	q	0566	ARMENIAN SMALL LETTER ZA	Glyphs nearly identical due to font design	Candidate
LATIN SMALL LETTER U	0075	u	u	057D	ARMENIAN SMALL LETTER SEH	Glyphs nearly identical due to font design	Candidate
LATIN SMALL LETTER IOTA	0269	ɩ	ʟ	0582	ARMENIAN SMALL LETTER YIWN	Glyphs nearly identical due to font design	Candidate

In-Script Variants

Preliminary Findings

- Until this moment no known “semantic” variants in researched languages
- One case of identical shapes: Turned E (01DD) and Schwa (0259)
- Certain cases under analysis:
  - Use of diacritics (e.g., shape and stacking)
  - IDNA 2003 Compatibility Issues (e.g., ‘ss’ and ‘ß’ (00DF))
  - Spacing (e.g., ‘æ’ (00E6) and ‘ae’)
  - Base character (e.g., ‘f’ (0066) and ‘ƒ’ (0192))

Other

WLE Rules Discussion

- No WLE Rules are planned for Latin LGR
- The contextual restrictions found in the Latin LGR proposal are for combining, non-space marks, e.g. U+0331 COMBINING MACRON BELOW
  - In the Latin LGR proposal, such marks are only allowed in defined sequences, i.e. only in combination with specific letter code points (in the code point order, only after a letter code point)
  - In the Latin LGR proposal, the letter code point and the combining mark are listed as a sequence in the LGR (in one instance, the sequence consists of a letter code point plus two combining marks)
  - By defining the combining marks in sequences only, no WLEs are needed for those
- No other code points need contextual restrictions

Test Labels File Discussion

- Test Label file has not yet been created

Questions?