
ICANN
COMMUNITY FORUM
58
COPENHAGEN
11–16 March 2017

Cyrillic GP Meeting

Dušan Stojičević | ICANN 58 | 13 March 2017

Agenda – Proposal for Cyrillic Script RZ LGR

1. Introduction and Background
2. Methodology
3. Results
4. Issues in MSR-2
5. Next Steps
6. IP Feedback and Discussion

Introduction – Script of LGR

- ISO 15924 Code: Cyrl
- ISO 15924 English Name: Cyrillic
- Latin transliteration of native script name: Cyrillic

- Maximal Starting Repertoire (MSR) version: MSR-2

Background on Script and Principle Languages

- Based on Early Cyrillic, from First Bulgarian Empire in 9th century AD
- Used for languages across Eastern Europe and north and central Asia
- Basis of alphabets in languages, past and present, especially those of Slavic origin, and non-Slavic languages influenced by Russian
- Used by more than 250 million people as the official script for their languages, about half from Russia

- With the accession of Bulgaria to the European Union in 2007, Cyrillic became **the third official script of the European Union, in addition to the Latin and Greek scripts**

Background on Script and Principle Languages

- Derived from Uncial script, augmented by letters from older Glagolitic alphabet, including some ligatures. Additional letters used for Old Church Slavonic sounds not found in Greek
- Named in honor of Byzantine brothers, Saints Cyril and Methodius, who created the Glagolitic alphabet
- Believed to be developed by disciples of Cyril and Methodius

- Individual languages and groups using Cyrillic script
  - Indo-European
  - Sino-Tibetan
  - Mongolian
  - Turkic
  - Individual - Aleutian, Nivkhs, Ket, Eskimos, Yukaghir languages
  - Caucasian
  - Chukchi and Kamchatka
  - Tungus
  - Ural

Geographic Territories Spread of Cyrillic

- Southeastern part of Europe (Serbia, Montenegro, Macedonia, Bulgaria, Bosnia and Herzegovina)
- Eastern Europe (Belorussia, Ukraine, Russia)
- Central Asia (Kazakhstan, Turkmenistan, Uzbekistan, Kyrgyzstan, Tajikistan, Mongolia)

is the only official orthography

is the only official orthography, but others are recognized for national or regional languages

Methodology

- According to work plan in the proposal for Cyrillic script GP
- Initially language based repertoire compiled, based on second-level IDN tables used by different ccTLDs, including the .su ccTLD which contained inventory for languages currently spoken in Russia
- Language repertoires collated in a face-to-face meeting in Istanbul on 25-26 Nov. 2016
- Continued to use the mailing list to share and finalize documents
- Consulted with Integration Panel (IP), including on crucial query regarding inclusion of U+02BC MSR for Ukrainian and Belarusian

Development Process – Inclusion Principle

1. Any code point which is a letter in established contemporary use in a language

Development Process – Exclusion Principles

1. Any code point DISALLOWED by IDNA 2008 protocol
2. Any code point representing a security or stability issue, which cannot be resolved at any other stage of the analysis (i.e., stage of determining code points, variants or whole label rules)
3. Any code point not listed in the MSR or listed in the MSR and deprecated or not recommended for use in Unicode Standard
4. Any code point representing technical signs only or that does not meet the inclusion criterion

Development Process – Exclusion Principles

5. The generation panel lacked sufficient information on the usage
6. The generation panel could only ascertain the use for such languages that had an EGIDS rating higher than five (6 or above), as per the “Guidelines for Developing Script-Specific Label Generation Rules for Integration into the Root Zone LGR”
7. 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

Code Point Repertoire

84 code points recommended for inclusion

8 code points recommended for exclusion (shown in the table)

#	Unicode CP	Glyph	Unicode Name	Lang. using CP	EGIDS value	Ref.
1	04EB	ӫ	CYRILLIC SMALL LETTER BARRED O WITH DIAERESIS	Khanty	Khanty 6b	Rule 6 http://www.omniglot.com/ writing/khanty.htm
2	04ED	ӭ	CYRILLIC SMALL LETTER E WITH DIAERESIS	Sami	Sami 8b	Rule 6
3	04DB	ӛ	CYRILLIC SMALL LETTER SCHWA WITH DIAERESIS	Khanty	Khanty 6b	Rule 6 http://www.omniglot.com/ writing/khanty.htm
4	04C2	ӂ	CYRILLIC SMALL LETTER ZHE WITH BREVE	Gagauz	Gagauz 5	Rule 5 http://www.omniglot.com/ writing/gagauz.htm Gagauz alphabet not in Cyrillic from 1996
5	04CC	ӌ	CYRILLIC SMALL LETTER KHAKASSIAN CHE	Khakas	Khakas 5	Rule 5 http://www.omniglot.com/ writing/khakas.htm
6	04CF	ӏ	CYRILLIC SMALL LETTER PALOCHKA	Khakas	Khakas 5	Rule 5 http://www.omniglot.com/ writing/khakas.htm
7	045D	ѝ	CYRILLIC SMALL LETTER I WITH GRAVE	Historical sign		Rule 6
8	0450	ѐ	CYRILLIC SMALL LETTER IE WITH GRAVE	Stressed sign		Rule 6

Cyrillic Script Variants

- No variants in Cyrillic script
  - Some code points visually confusable
    - not considered as variants by the Cyrillic community
    - provide table of confusable code points, so organizations can use as needed

Cross-Script Variants

- Decided to limit these to homoglyphs
- Included code points which are homoglyphs in the lower case but not homoglyphs in the upper case
  - Only lower case because upper case disallowed in IDNA 2008
  - Decision made in consultation with IP (“the IP, at this point, does not require that upper case homoglyphs are included”)
- Cyrillic GP found cross-script variants with:
  - Armenian
  - Greek
  - Latin
- Cyrillic GP did not find cross-script variants with Georgian

Cross-Script Variants – with Armenian Script

- Armenian GP indicates three (3) variants with Cyrillic script
- Opinion of Cyrillic GP that only one (1) homoglyphic variant
- Other two (2) not identical, so included in confusables table

Armenian glyph	Armenian code point	Cyrillic glyph	Cyrillic code point
օ	0585	о	043E

Cross-Script Variants – with Greek Script

- Cyrillic has three (3) homoglyphic variants with Greek script

Greek glyph	Greek code point	Cyrillic glyph	Cyrillic code point
κ	03BA	к	043A
ο	03BF	о	043E
ϕ	03C6	ф	0444

Cross-Script Variants – with Latin Script

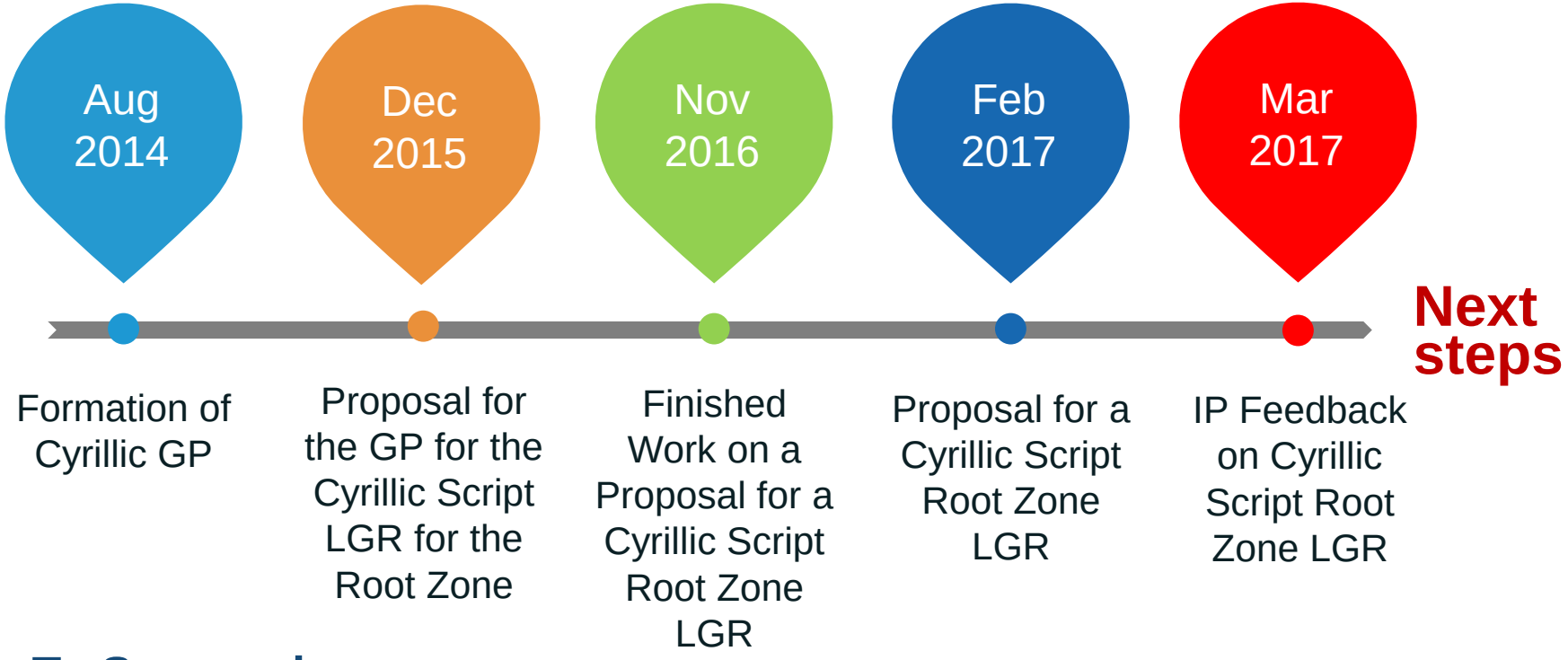
- Cyrillic has following homoglyphic variants with Latin from MSR-2
  - Confusables listed separately

Latin glyph	Latin code point	Cyrillic glyph	Cyrillic code point
a	0061	а	0430
c	0063	с	0441
e	0065	е	0435
o	006F	о	043E
i	0069	і	0456
j	006A	ј	0458
l	006C	ӏ	04CF
p	0070	р	0440
s	0073	ѕ	0455
y	0078	у	0443
x	0079	х	0445
ä	00E4	ӓ	04D3
ë	00EB	ё	0451
æ	00E6	ӕ	04D5
ə	01DD	ә	04D9

Discussion of Issues in MSR-2

- Ukrainian and Belarusian languages have “apostrophe” (U+02BC) as a letter – not punctuation sign
  - *IP response*
- Montenegro has two new chars in national scripts (Latin and Cyrillic); ccTLD to implement them at second level
  - *Not yet in Unicode*
- Old Church Cyrillic script
  - *Not in second level use, it should not be used for root zone*
- Using upper case and lower case Unicode code points in Cyrillic as problem of the confusion during visualization
  - *Need to address that*

Timeline



To Summarize

It took Cyrillic GP two and a half years to get to the stage with final proposal. But, the work has been done according to the dates defined in original working plan.

Next Steps

1. Short phase (planned for this session)
   i. Finalize proposal, based on IP feedback
   ii. Finalize XML and test data files
   iii. Issue the LGR for public comment

2. After public comment phase
   iv. Finalize the LGR proposal to include community feedback

3. Long-term phase
   v. Address new code points included in the MSR in the future
      - if needed in the root zone LGR, GP to re-convene and create additional proposal

Members of Cyrillic GP

Alex Khmyl (Belarus)	Nelly Stoyanova (Bulgaria)
Alexei Sozonov (Russia)	Nodir Mirzoev (Tajikistan)
Almaz Bakenov (Kyrgyz Republic)	Oleksandr Tsaruk (Ukraine)
Daniel Kalchev (Bulgaria)	Pavel Gusev (Kazakhstan)
Dmitry Belyavskiy (Russia)	Predrag Lesic (Montenegro)
Dmitry Kohmanyuk (Ukraine)	Sanja Simonova (Macedonia)
Dušan Stojičević (Serbia, chair)	**Sergey Povalishev (Belarus)**
Enkhbold Gombo (Mongolia)	Tattu Mambetalieva (Kyrgyzstan)
Iliya Bazlyankov (Bulgaria)	Yashar Hajiyev (Azerbaijan)
Kadamjon Safiev (Tajikistan)	Yuliya Morenets (Russia)
Mirjana Tasić (Serbia)	**Yurii Kargapolov (Ukraine)**
Nazgul Kurmanalieva (Kyrgyzstan)	Yuriy Honcharuk (Ukraine)

IP Feedback

- Procedural point on value of incomplete submission
  - IP needs machine-readable LGR and test labels including invalid labels, to use tools for reviewing proposals, such as mechanically verifying that they are proper subsets of the MSR, or comparison against other data sets
  - With no variants other than cross-script homoglyphs and no WLE rules, the IP reverse-engineered XML for analysis
  - File shared with Cyrillic GP who may use it in any way to assist the GP in creating an XML to accompany next submission
  - Generation Panels consider no draft "complete" without formal specification of LGR according to RFC 7940

IP Feedback

- Other procedural points on style of submission

1. References in the main repertoire table be numbered, e.g. [106] to match the XML - sample XML prepared by IP

2. IP submitted XML file to the GP, as example satisfying RFC 7940 and IP's formatting requirements - GP could review and complete

3. HTML version also attached – generated mechanically from XML

IP Feedback

- Other substantive points on style of submission

1. Section 5.4 “Code points excluded” refers to “Rules”. “Exclusion principles” in Section 5.2? Should consistently refer to as “principles” or “Exclusion principles”

2. Preferable if GP state explicitly which languages it reviewed, and which are, finally, supported by LGR

IP Feedback

- Other substantive points on style of submission

1. In point 6 of Section 5.2, exclusion if a language has EGIDS rating higher than 5 (i.e. 6 or above).
   - Implies languages with EGIDS 5 or below included
   - Root Zone LGR to cover score of 4 and below
   - For languages with EGIDS 5, review needed to determine support in LGR
   - Not clear which languages considered and which finally supported
   - When supporting such border-line (EGIDS-5) languages, expect reasoning behind decision, with citation of evidence

2. If code point for a language of EGIDS 4 excluded, expect to be discussed with specific details

IP Feedback

- Specific inadequacies or obscurities in proposed repertoire
  - In Section 5.4, the status of the following code points from MSR-2 are not substantiated
    1. 04ED SMALL LETTER E WITH DIAERESIS : No reference is supplied. Perhaps the LGR could refer to https://en.wikipedia.org/wiki/Kildin_Sami_orthography

       04C2 SMALL LETTER ZHE WITH BREVE: Substantiated with “Rule 5” (presumably: Principle 5 – lack of sufficient information). However, the note states that the language is no longer written in Cyrillic (with a time-out in 1996)

       - Included as optional code point (“for extended use”) in Reference Second Level LGR for Ukrainian. Good if Cyrillic GP could comment to resolve any perceived differences

IP Feedback

- In Section 5.4, the status of the following code points from MSR-2 are not substantiated

  2. 04CF SMALL LETTER PALOCHKA: Excluded but widely used in provincial and educational languages. For example, Wikipedia [Palochka] notes used in several languages with EGIDS 4 or smaller, for example Adyghe 2, Chechen 2, Avar 3, Dargwa 4, Ingush 4, Lak 4, Lezgian 4, Tabassaran 4, as well as Abaza 5, and perhaps Kabardian 5.

     - GP give detailed account of reasoning behind decision to exclude 04CF

IP Feedback

- In Section 5.4, the status of the following code points from MSR-2 are not substantiated

    3. Code point from MSR-2 neither included nor listed as excluded

       0525 CYRILLIC SMALL LETTER PE WITH DESCENDER

       (per Unicode encoded for Abkhaz, included language)

IP Feedback

- In Section 5.4, the status of the following code points from MSR-2 are not substantiated

4. Following code points omitted
   0450 è CYRILLIC SMALL LETTER IE WITH GRAVE
   045D ѝ CYRILLIC SMALL LETTER I WITH GRAVE

   - In reference second-level LGRs - for Bulgarian and Macedonian
   - For Bulgarian only marginal – i.e. “available for extended use”
   - For Macedonian, requested for addition to the Unicode with evidence for their use: http://www.unicode.org/wg2/docs/n1323.pdf (10MB)
   - May nevertheless be valid reason for treating differently in Root Zone LGR. If so, differences be described in LGR proposal
   - decision to exclude these code points be briefly described

IP Feedback

- In Section 5.4, the status of the following code points from MSR-2 are not substantiated

    4. 04C2 ӂ CYRILLIC SMALL LETTER ZHE WITH BREVE
    The third, 04C2, is used apparently only in two languages:

        - Gagauz and Moldovan. Gagauz (EGIDS 5) is primarily spoken in Moldova, but uses Cyrillic only in Russia, Ukraine and Kazakhstan;

        - Moldovan, though official in Moldova (hence EGIDS 1), uses Cyrillic characters in Transnistria only.

IP Feedback

- Armenian homoglyphs
  - Not include as cross-script homoglyphs mapped in the Armenian LGR
    - 0448 CYRILLIC SMALL LETTER SHA
    - 04BB CYRILLIC SMALL LETTER SHHA Based on [Procedure] the integrated LGR
  - RZ LGR will contain union of variants defined in the individual LGRs

- Other homoglyphs
  - Section 6.2.4 lists 04CF SMALL LETTER PALOCHKA, though not in repertoire

Cyrillic Script Root Zone LGR: IP Feedback

- Conclusion
  - IP's general impression that well-constructed LGR, a serious contribution to managing use of vast Cyrillic alphabet within the Root Zone
  - IP requests GP to provide the missing files and some of the essential rationale as itemized

Thanks!

Dušan Stojičević
Душан Стојичевић

dusan@dukes.in.rs
stojicevic@gransy.com

Engage with ICANN and IDN Program

Thank You and Questions

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

twitter.com/icann

soundcloud.com/icann

facebook.com/icannorg

weibo.com/ICANNorg

youtube.com/user/icannnews

flickr.com/photos/icann

linkedin.com/company/icann

slideshare.net/icannpresentations