Joint Meeting of Latin GP and IP at ICANN66

Latin GP Members

Agenda Overview

1 Short History

2 Scope of Work

3 Members

4 Work Accomplished

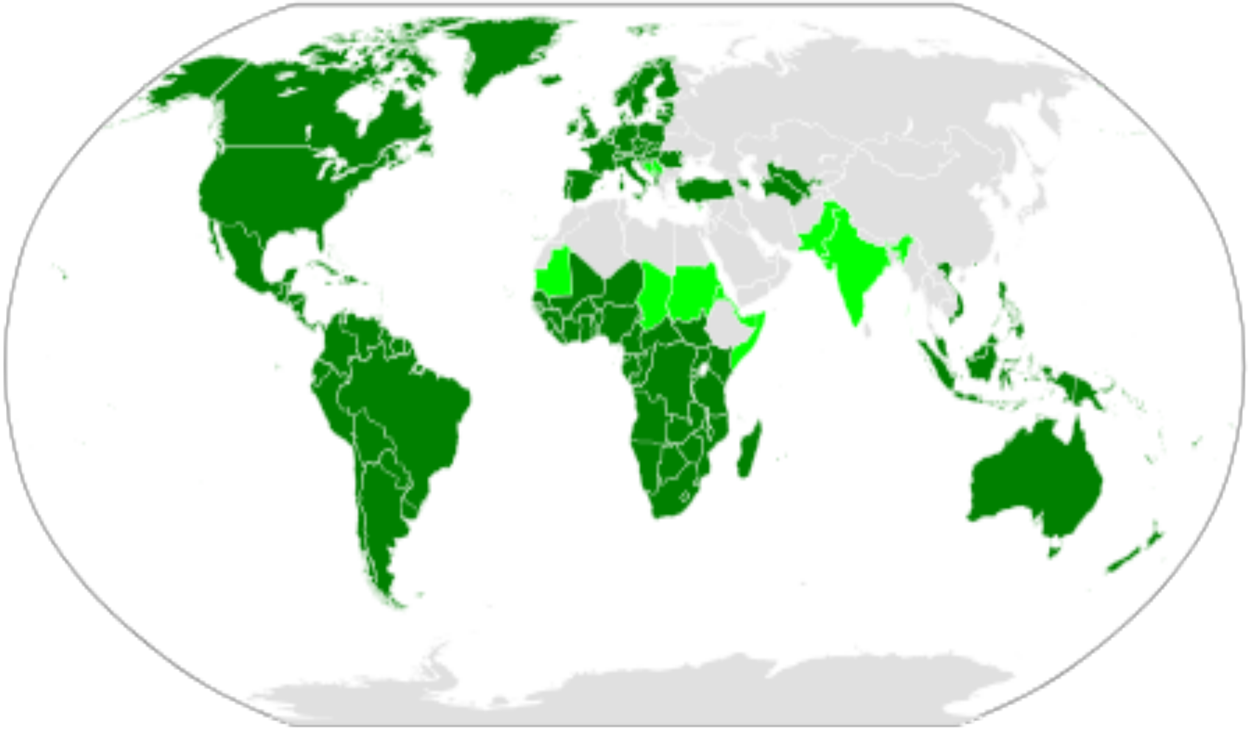
5 Project Timeline

6 In-Script Variant Analysis

Latin GP – Short History

- June-August 2016 - GP restarted with a new call for volunteers.
- 15 May 2017 - The Latin GP is seated.
- September 2017 - GP proposal for Principals on inclusion and exclusion of code points was sent for an informal public review.
- September-November 2017 - GP collected information on 455 languages that use Latin script.
- May 2018 (for MSR-3) and October 2018 (for MSR-4) - GP submitted the code point repertoire to the Integration Panel.
- May 2018 – GP submitted the updated LGR proposal with repertoire.
- January 2019 - GP submitted the updated LGR proposal with the cross-script variant analysis and the initial in-script variant analysis.
- October 2019 - GP submitted the updated LGR proposal with some in-script variant analysis.

Latin Script Geographic and Linguistic Spread



Dark green = Latin script is the main script.

Light green = Latin co-exists with other scripts.

Grey = Latin-script alphabets are sometimes extensively used 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 in China or in Japan.

Latin GP – Scope of Work for Code Point Analysis

- Maximal String Repertoire Version 4 (MSR-4)
  - Subset of code points allowed in 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.
- Proposed repertoire has 210 Latin code points.

Latin GP – Scope of Work for Variant Analysis

- In-script variant analysis
  - Visual variants
  - Non-visual variants

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

- Other considerations:
  - Basic shapes (e.g., circle “o”, single line “l”, and crescent “c” or “ɔ”) within all scripts.
  - Underlining analysis
  - IDNA2003 compatibility

Latin GP – Members

- 14 members (7 active members), 3 observers
- Language representatives
  - Africa
  - Asia
  - Australia and Oceania
  - Europe
  - North America
- Diversity
  - Community representatives
  - Linguistic experts
  - Registry/registrar experts
  - Technical community, DNS experts
  - IDNA/Unicode experts

Latin GP – Challenges and Solutions

- Challenges

  Many languages

  Multiple code points to process

  Change of requirements

  Complex in-script variant analysis

- Solutions

  Process languages with EGIDS=1-4 first (181).

  Consider processing languages with EGIDS=5 (110).

  - 29 languages with at least one million users with sufficient reference are included.

  Define simple procedure for developing Latin script repertoire.

  Workload divided in two groups:

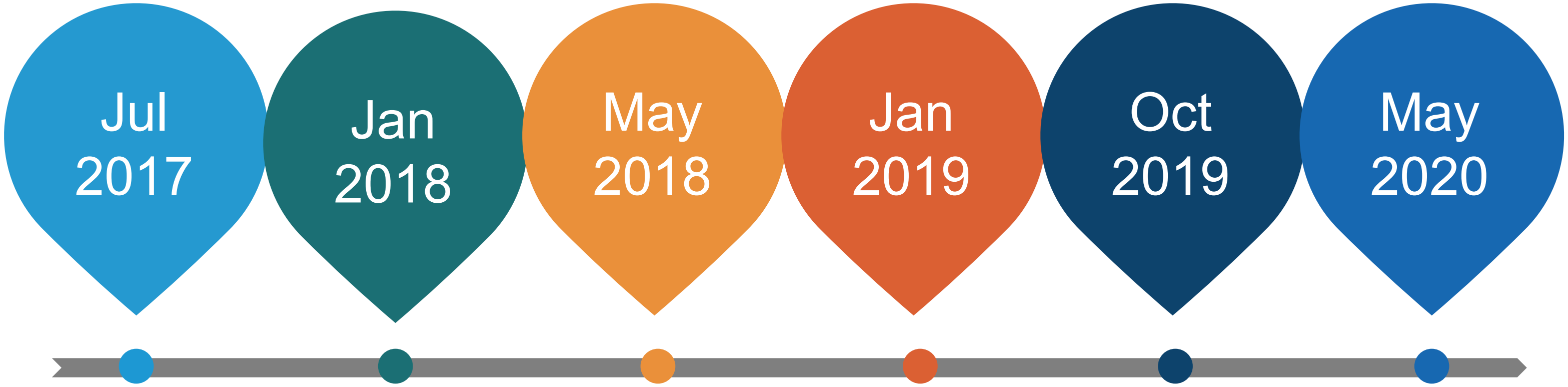
  - Repertoire Working Group
  - Variant Working Group

  Extend planned working time (finish 2020 instead of 2018).

Latin GP – Work Accomplished

- Developing Repertoire
  - 181 of 181 EGIDS 1- 4 languages processed.
  - 29 EGIDS 5 languages processed.
  - 193 of 279 MSR-2 code points attested.
  - 3 non-MSR-2 code points are included in MSR-3.
  - 3 non-MSR-3 code points are included in MSR-4.
  - 22 Code Point Sequences identified.
- Developing Variants
  - In-script variants still ongoing (80% finished).
  - Cross-script variants with Armenian script defined.
  - Cross-script variants with Cyrillic script defined.
  - Cross-script variants with Greek script defined.
  - Special HTML Link (underlining) analysis completed.
  - IDNA2003 compatibility analysis completed.
- Submitted the third version of proposal to the IP in May 2018.
- Submitted the fourth version of proposal to the IP in Jan 2019.
- Submitted the fifth version of proposal to the IP in Oct 2019.

Latin GP – Project Timeline



Jul 2017
Developed Principles

Jan 2018
Developed code points repertoire and identified variants.

May 2018
Submitted proposal with code points repertoire.

Jan 2019
Submitted proposal adding cross-script variant analysis.

Oct 2019
Submit proposal without complete in-script variant analysis.

May 2020
Finalize the LGR proposal and publish for Public Comment.

Latin GP – In-Script Variant Analysis

1 Work organization

2 Definition of In-script variant analysis framework

3 Visual Variants

4 Underlining

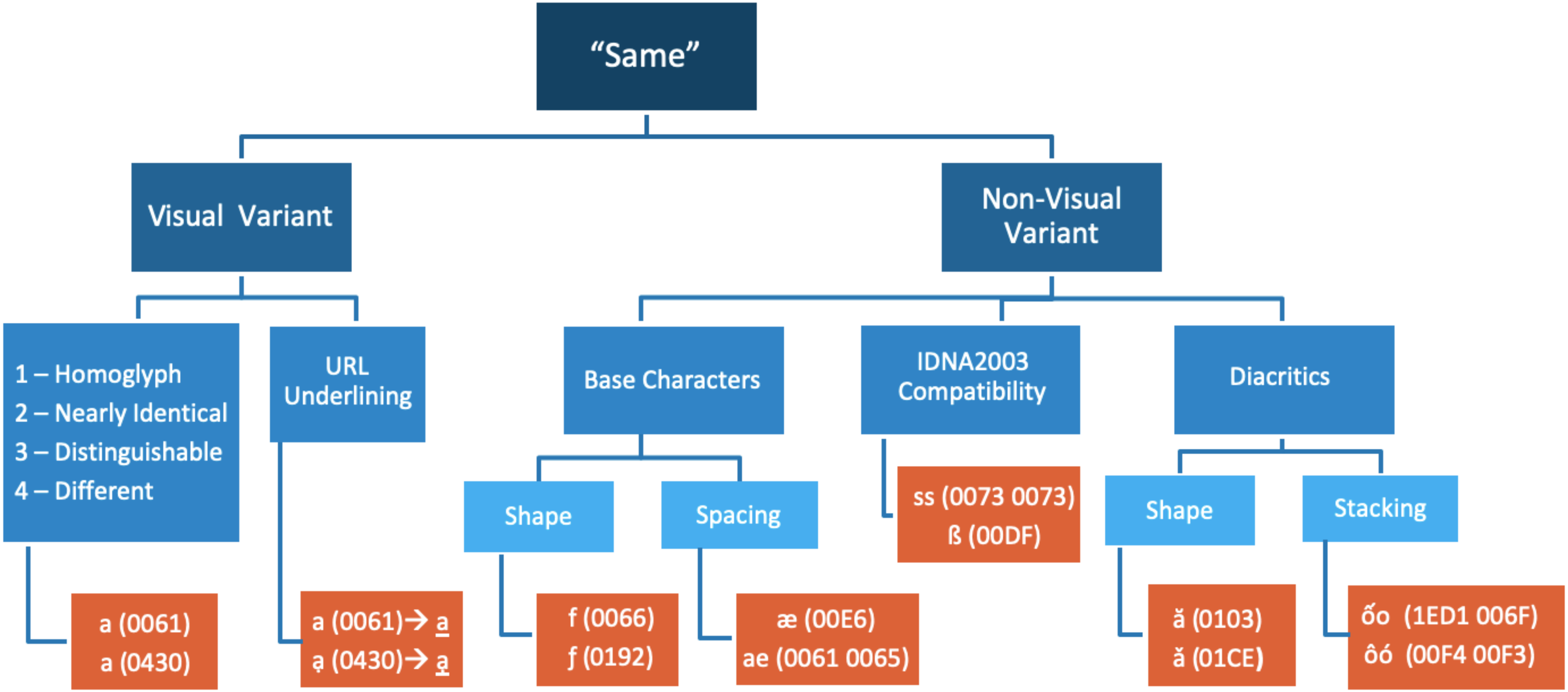
5 Non-Visual variants

6 IDNA2003 problem

Latin GP – In-Script Variant Analysis – Work Organization

- All investigations of code points should be done using wordmark.it.
- Only Google fonts should be taken into consideration.
- All seven members get the same amount of code points to analyze.
- Results of analysis should be presented in the proposed template.
- All results of analysis should be discussed by GP before accepting code points as an in-script variants.
- All working material will be presented in the Final Report as Appendix.

Latin GP - In-Script Variant Analysis Framework



Latin GP – Variant Principles Matrix Proposal

(to be finalized)

Index #	Principle	Reason	Disposition
1	Visual variant (homoglyph)	Security	Blocked
2	Visual variant (glyph nearly identical)	Security	Blocked
3	Visual variant (generally acceptable font design)	Security	Blocked
4	Non-visual variant	Security	Blocked
5	Symmetry property {a:b}	Security	Blocked
6	Transitivity property {a:b; b:c}	Security	Blocked
7	URL underlining	Security	Blocked
8	IDNA2003 Compatibility	Security	Blocked
9	Function (alternate orthography)	Usability	Allocatable

Latin GP – In-Script Variant – Visual Analysis

Score	Category
1	**Homoglyphs** A pair of code points in this category has essentially identical appearance by design.
2	**Nearly Identical** A pair of code points is considered Nearly Identical when the visual confusion can be attributed to font design.
3	**Distinguishable** A pair of code points is considered Distinguishable when any of the code point's glyphs have recognizably different features from the other code point.
4	**Different** When the two glyphs in the pair are sufficiently different.

Latin GP – In-Script Variant – HTML Underlining

- In many browsers and word processing software tools, links to websites are indicated by underlining the domain name.

- This has obvious implications when the codepoint involves a diacritic below the line.

- In some cases, the diacritic is entirely or partially obscured by the underline.

- In other cases, the underline merely makes it very difficult to discern just what is going on below the line.

- To address this, Latin GP has investigated both:

    The code points themselves.

    The code points when underlined.

Latin GP – In-Script Variant – HTML Underlining

Group	Underlining							
Target			**Source**			**Variant Candidate [Yes/No]**	**Disposition [Allocatable / Blocked]**	**Rationale**
Code Point	**Glyph**	**Name**	**Code Point**	**Glyph**	**Name**			
0061	a	Latin Small Letter A	0105	ą	Latin Small Letter A With Ogonek	YES	Blocked	Glyphs nearly identical due to underlining
0061	a	Latin Small Letter A	0061 + 0331	a	Latin Small Letter A + Combining Macron Below	YES	Blocked	Glyphs nearly identical due to underlining
0061	a	Latin Small Letter A	1EA1	a	Latin Small Letter A With Dot Below	YES	Blocked	Glyphs nearly identical due to underlining

Latin GP – In-Script Variant – Non-Visual Analysis

- **Shape of Base Characters** - Latin GP hypothesized that some hand-written forms may end up taking similar or the same shapes as some derived letters, and that readers may consider such unknown derived letters as hand-written variations of familiar letters, such as e.g. **v vs. ʋ**.

- **Spacing of base characters -** Several letters have been derived by putting more closely together sequences of two or more glyphs. Depending on the spacing in between those glyphs in a font, ligatures may become indistinguishable from a sequence of glyphs of which the same ligature is composed e.g **ae** vs **æ**.

Latin GP – In-Script Variant – Diacritics

- **Shaping of diacritics** - three types of potential issues with diacritic modifiers:

  Certain diacritics may be considered to be conceptually the same as others by significant parts of the user community, such as **dot below or a comma below**.

  Certain diacritics are not kept apart from one another in handwriting traditions:

  - **caron** being written in the same way as a **breve**.
  - **dot above** being written in the same way as an **acute**.
  - **diaeresis** being replaced by **two vertical strokes** could be mistaken for a **double acute** in italic fonts.
  - **tilde being** written ‘simply’ as a simple horizontal stroke above, i.e. a **macron**.

Latin GP – In-Script Variant – Diacritics

- **Shaping of diacritics** - three types of potential issues with such modifiers:
  - Number of diacritics are only used in a very limited part of the script using community. This may lead to confusion:
    - **Horn** (ʼ) could be conceptually mistaken by some readers for a misplaced **acute** (´) or even an **apostrophe** (‘)
  - For someone to discern a difference between code points or between diacritics, they have to be aware that they exist. The average user, and even the "very careful user," is familiar with perhaps a half dozen of diacritics. As a result, they do not realize that it would be advisable to look for differences.

Latin GP – In-Script Variant – Diacritics

- **Stacking of diacritics** - diacritics are combined with one another, such as **ấ** featuring both a circumflex and an acute. Glyphs featuring base characters with several diacritics:

    May become visually identical or confusable to readers with sequences of glyphs featuring the same diacritics on two separate code points.

    May even become effectively invisible in context by lapping over into adjacent glyphs.

- **Combining diacritics**

    Does not combine in some fonts (e.g. Courier New).

    In some fonts, it combines with the next code point and does not stay on the correct code point.

Latin GP – In-Script Variant – IDNA Compatibility

- **00DF (LATIN SMALL LETTER SHARP S)** – IP expected from Latin GP “to consider the inclusion of this code point in the LGR and to investigate the case for or against making it a blocked variant of **ss**.”

- **0069 (LATIN SMALL LETTER I) and 0131 (LATIN SMALL LETTER DOTLESS I)** - case operations are locale sensitive. IP expected from Latin GP “to investigate the need to address any compatibility issues related to this code point, and if found, to suggest means to mitigate them.”

Questions?