Proposal for Generation Panel

for Latin Script Label Generation Ruleset for the Root Zone

# General Information

The Latin script[[1]](#footnote-1) or Roman script is a major writing system of the world today, and the most widely used in terms of number of languages and number of speakers, with ca. 70% of the world’s readers and writers making use of this script[[2]](#footnote-2) ([Wikipedia](http://www.webcitation.org/6oGZwoNUu)).

Historically, it is derived from the Greek alphabet, as is the Cyrillic script. The Greek alphabet is in turn derived from the Phoenician alphabet which dates back to the mid-11th century BC and is itself based on older scripts. This explains why Latin, Cyrillic and Greek share some letters, which may become relevant to the ruleset in the form of cross-script variants.

The Latin alphabet itself originated in Italy in the 7th Century BC. The original alphabet contained 21 upper case only letters, i.e. A, B, C, D, E, F, Z, H, I, K, L, M, N, O, P, Q, R, S, T, V and X. Soon after, this repertoire was extended, and e.g. letter G developed from C and J from I, while letter V and U split and a ligature[[3]](#footnote-3) of VV became W. The basic set of the historic repertoire of Latin script is considered: A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y and Z, and Latin script is generally written left-to-right.

With the spread of the Roman Empire and the Catholic Church, the script became increasingly used to represent other languages apart from Latin, and so new letters were added to the script using different mechanisms for the purposes of representing speech sounds which were unknown to - at first - Latin or - much later - European languages: For example, some letters were formed by combining two glyphs into one ligature, such as 'æ' from 'a' and 'e', as used in Danish and Norwegian. Also, new letters were borrowed through contact in between languages and scripts, such as 'þ' (thorn) used for Scandinavian languages, which was borrowed from the Runic alphabet, or 'Ƹ/ƹ' borrowed from the Arabic letter ' ع' ([Wikipedia](http://www.webcitation.org/6oGjj2nLB)) and borrowing was an ongoing process, which continued throughout the history of the use of Latin Script. Even repeated processes of borrowing occurred, such as e.g. Greek 'Ɣ' (gamma), which was the original source for both C and G, but which was re-borrowed again in it’s original shape for the representation of the voiced velar fricative for the writing of languages such as Ewe (ISO 639-3[[4]](#footnote-4) ewe) ([Wikipedia](http://www.webcitation.org/6oGbPb0E5)).

Apart from borrowing, new letters were developed using different mechanisms such as the addition of various modifiers to existing letter shapes, such as bars or strokes, e.g. 'Ɍ/ɍ' as used in Kanuri (ISO 639-3 kau) ([Wikipedia](http://www.webcitation.org/6oGhlKnWo)), hooks, e.g. 'Ƙ/ƙ' as used in Hausa (ISO 639-3 hau) ([Wikipedia](http://www.webcitation.org/6oGiAslOh)), horizontal rotation or mirroring, e.g. 'Ǝ' based on 'E' as used by the Pan-Nigerian alphabet ([Wikipedia](http://www.webcitation.org/6oGj088MG)), vertical rotation or mirroring, e.g. 'Λ/ʌ' on the basis of 'v', as used by Ibibio (ISO 639-3 ibb) ([Wikipedia](http://www.webcitation.org/6oGkNfVVD)), making (part of) a letter shape more cursive or italic, e.g. 'Ʋ/ʋ' on the basis of 'v' as used by Ewe (ISO 639-3 ewe) ([Wikipedia](http://www.webcitation.org/6oGbPb0E5)), or the re-use of further signs such as punctuation marks to develop letters, e.g. 'Ɂ' (glottal stop) based on '?' (question mark) as used by Chipewyan (ISO 639-3 chp) ([Wikipedia](http://www.webcitation.org/6oGkjW05y)).

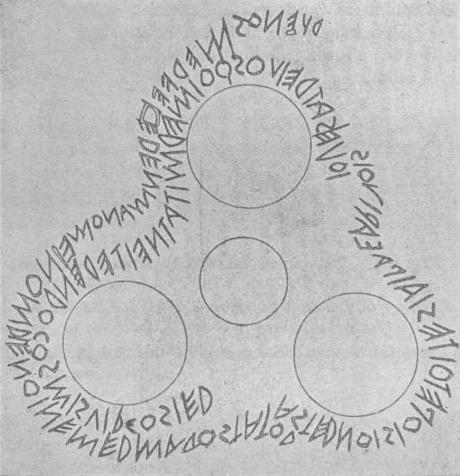
Typologically the Latin script is an Alphabet, which means that segments or units of the writing system generally tend to represent consonants and vowels (rather than other linguistic units)[[5]](#footnote-5). Today, Latin script is a bicameral system, which features upper and lower case forms for most letters, but there may be little visual similarity between a letter’s upper and lower case forms, for example, 'A' and 'a'. However some orthographies have reversed the development back towards a unicameral system, without a distinction in between upper and lower case letters, such as the second version of the African Reference Alphabet from 1982 ([Wikipedia](http://www.webcitation.org/6oGdYWz8B)).

Figure 1: The Duenos Inscription, 6th Century B.C., one of the earliest surviving documents in Latin (taken from [Wikipedia](https://commons.wikimedia.org/wiki/File:Duenos_inscription.jpg))

## 

## 1.1 Use of Latin Script characters in domain names

Traditional domain names are domain names without the extension that IDN gives, i.e. the domain names that most of us are familiar with. The characters in traditional domain names are limited to a subset of the ASCII characters set. In Top Level Domains (TLDs) the permitted characters are the characters in the English alphabet, i.e.:

a b c d e f g h i j k l m n o p q r s t u v x y z

Those characters, 'a-z', is a subset of the Latin Script.

In traditional domain names, or traditional domain name labels, both upper case 'A-Z' and lower case 'a-z' can be used and mixed, and are considered to be equal. E.g. “com”, “Com” and “COM” are identical from a domain name perspective. We will see below, that in IDN domain name labels, only the lower case 'a-z' may be used.

However, in Second Level Domains (SLDs) and below, digits (0 1 2 3 4 5 6 7 8 9) and the hyphen '-' can also be included. As in all domain names, the dot '.' is used to separate the parts (or “labels”) of the domain name, e.g. “www”, “example” and “com” in “www.example.com”. In this example “com” is the traditional TLD label that may only contain characters 'a-z', whereas the other label may also contain digits and the hyphen.

## 1.2 Target Script for the Proposed Generation Panel

As per the *Procedure to Develop and Maintain the Label Generation Rules for the DNS Root Zone in Respect of IDNA Labels* (referred to simply as [Procedure] in the following), only code points included in Maximal Starting Repertoire, Second Version (Referred to simply as [MSR-2] in the following) will be considered.

The Latin script has the following specifications:

ISO 15924 code: Latn

ISO 15924 no.: 215

English Name: Latin

The set of code points in the Latin script, as specified by [MSR-2], contains the following Unicode ranges:

|  |  |
| --- | --- |
| Script | Range of Unicode code points |
| Controls and Basic Latin | U+0061 – U+007A |
| Controls and Latin-1 Supplement | U+0080 – U+00FF |
| Latin Extended-A | U+0100 – U+017F |
| Latin Extended-B | U+0180 – U+024F |
| IPA Extensions | U+0250 – U+02AF |
| Combining Diacritical Marks | U+0300 – U+036F |
| Combining Diacritical Marks Supplement | U+1DC0 – U+1DFF |
| Latin Extended Additional | U+1E00 – U+1EFF |
| Latin Extended-C | U+2C60 – U+2C7F |

MSR-2 excluded the following Latin script UNICODE ranges:

|  |  |
| --- | --- |
| Latin Extended-D | U+A720 – U+A7FF |
| Latin Ligatures | U+FB00 – U+FB0F |
| Full-width Latin Letters | U+FF00 – U+FF5E |

* Latin Extended-D; technical use (phonetic)/obsolete/punctuation
* Latin Ligatures; compatibility characters not PVALID in IDNA 2008
* Full-width Latin letters; compatibility characters not PVALID in IDNA 2008

Furthermore, only lower case letters are considered, as upper case ones may not be used in IDNs following [IDNA 2008] as well as earlier traditions of DNS (cf. [RFC1034], [RFC1035]) and are excluded from [MSR-2].

### 1.2.1 Diacritics

Diacritics are marks which combine together with other letters. In Latin script, many languages make use of diacritics to modify letters. Diacritics may appear anywhere around, most commonly above 'é', below 'ṯ', or through 'Đ' a letter. Several diacritics may attach to the same letter, such as in Vietnamese, e.g. 'ệ'. However, often diacritics are used for specialized purposes, such as phonetic notation/romanization and therefore may not be part of the general orthographic system.

Diacritics may perform different roles depending on the language: Some languages consider letter + diacritic as one letter, such as Norwegian (both Bokmål and Nynorsk varieties), for example, which lists 'Æ', 'Ø' and 'Å' at the end of its alphabet. Meanwhile in Italian, diacritics are recognized as a further layer of the orthography, and for example 'è' which differs from 'e', is not listed as a separate letter of the alphabet, but occurs in a dictionaries following unmodified 'e'.

Similarly diacritics may express different units of of language and may be treated differently from other elements of the writing system. In numerous languages they are obligatory part of the alphabet and serve to distinguish entirely different phonemic units of a language, such as the combining stroke[[6]](#footnote-6) on 'l' in Navajo, i.e. 'l' vs. 'ł', where the former represents an approximant and the latter a fricative, two different types of consonants. Also it diacritics may be applied to other letters to systematically modify phonemes, such as U+0308 ̈ (COMBINING DIAERESIS) which in German is an obligatory part of the alphabet generally indicating a change in the quality of a vowel as in "*Ofen"* [ˈoːfən] "oven" vs. "Öfen" [ˈøːfən], the plural form of the same. However in other languages, they may be optional in some contexts, such as in Italian, where homographs, which differ in the position of the word-level stress (the so-called *parole sdrucciole)*, can optionally be written with an combining accent mark U+02CA 'ˊ' (MODIFIER LETTER ACUTE ACCENT) to indicate the position of word-level stress in antepenultimate positions, .e.g. "*súbito"* "immediately" vs. "*subito"* "suffered" ([about.com](http://italian.about.com/od/pronunciation/fl/italian-accent-marks.htm)). Diacritics may express segmental units of a languages such as single phonemes, or it may express supra-segmental features such as word-level stress, e.g. such as U+02CA 'ˊ' (MODIFIER LETTER ACUTE ACCENT) in Spanish "*corazón"* "heart", or tone such as in Thai-word "*mái"* "wood", where the same indicates a high tone. Nonetheless diacritics may serve to distinguish minimal pairs in numerous languages, irrespective of whether their omission is considered a spelling mistake, such as "*vô"* "grandfather" vs. "*vó"* "grandmother" in Portuguese or "*è"* "is" vs "*e"* "and", in Italian, or whether the use of such diacritics is optional, such as in the Italian example "*súbito"* vs. "*subito"* presented above.

While generally diacritics take effect only on those letters with which they are combined, there are few languages in which diacritics have an effect on adjacent letters, particularly where several letters are used to represent single phonemes, i.e. di-, tri- or quadri-graphs, for example in Maltese where the combining stroke over 'h' affects the preceding letter 'g', since 'għ' is a digraph in that language which contrasts with simple 'g', e.g. "*gallettina"* "biscuit" and "*għasfur"* "bird" ([Wikipedia](http://www.webcitation.org/6oIqqFqSd)).

Nonetheless diacritics may be omitted in the context of IDNs as they are not always necessary for word recognition by the relevant language communities even where the omission of such diacritics constitutes a spelling mistake, such as "café", which must be spelled with the accent in both German and English, but which is represented as "cafe" in numerous labels of the .de ccTLD zone.

Technically, most diacritics[[7]](#footnote-7) are encoded as combining marks in Unicode and a letter with two diacritics such as 'ḉ' may be represented in several ways in Unicode – as a pre-composed form (U+1E09), or as any one out of a number of different possible combinations of several code points, such as the letter and the first diacritic with the second added (U+0107 'ć' + U+0327 ̧ COMBINING CEDILLA). However [IDNA 2008] specifies that U-labels must be in Unicode *Normalization Form C* (NFC), the pre-composed form U+1E09 in the given example, as defined by Unicode - cf. [Unicode Standard Annex #15].

Accordingly, these multiple representations are usually eliminated by normalization, except for cases where no precomposed forms have not yet been encoded in Unicode 6.3 (on which [IDNA 2008] is based). One example is 'ɔ̱', i.e. a sequence of U+0254 (LATIN SMALL LETTER OPEN O) and U+0331 (COMBINING MACRON BELOW) (or in reversed order), as used by the orthography of Nuer (ISO 639-3 nus) ([Scriptsource](http://www.webcitation.org/6oJkOD7S5)), where it indicates a breathy open-mid back rounded vowel 'ɔ̤'.

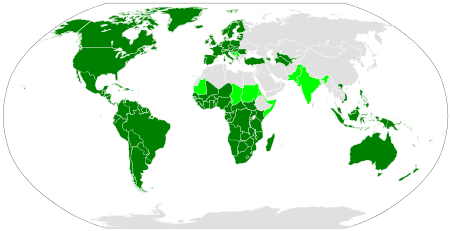
### 1.2.2 Latin Script as Represented in Unicode and Variant Relationships With Other Scripts

As represented in Unicode, the Latin script has some identical glyphs, for example, 0259 'ə' (LATIN SMALL LETTER SCHWA) and 01DD 'ǝ' (LATIN SMALL LETTER TURNED E), which need to be handled by variant rules. However, there may also be variants occurring across scripts with both related as well as unrelated scripts: Since Latin and Cyrillic scripts developed from the Greek script they share some letters[[8]](#footnote-8). The Armenian script may be modelled on the Greek script and a small number of letters are shared. For example, Cyrillic script shares a number of glyphs with Latin, including 0430 'а' (CYRILLIC SMALL LETTER A), 0435 'е' (CYRILLIC SMALL LETTER IE), 0455 'ѕ' (CYRILLIC SMALL LETTER DZE), 0458 'ј' (CYRILLIC SMALL LETTER JE). . However quasi homoglyphs may also exist in between Latin script and other scripts, such as Greek 03BF '*ο'* (GREEK SMALL LETTER OMICRON) and Latin 006F 'o' (LATIN SMALL LETTER O) ([Wikipedia](http://www.webcitation.org/6oaXaUdrR)). Homoglyphs may also occur with less closely related or unrelated scripts, and Latin GP will particularly consider results from previous generation panels, such as Armenian GP which has identified 7 codepoints which they considered as blocked variants [Armenian Proposal: 5]. Additional homoglyphs in between scripts may occur which however may effectively become irrelevant at the root zone level due to the Letter Principle in the [Procedure], which states that only code points exclusively used in writing words are to be included in root zone labels. For example, 006C 'l' ( LATIN SMALL LETTER L) and Arabic script numeral 0661 '١' (ARABIC-INDIC DIGIT ONE) could be considered quasi homoglyphs depending on rendering of the user's client, but since the Arabic script codepoint is a number, it remains irrelevant to the root zone (but may become relevant at further levels).

Similarly, under-resourced or minority languages may make use of code-points which not yet been encoded in Unicode, or which have been encoded revisions subsequent to Unicode 6.3 (on which [IDNA 2008] and [MSR-2] are based). Also combining marks may be required for some languages eligible for inclusion in the LGR, and the panel may discover evidence of the use of code points in orthographies of languages eligible for inclusion, which are combining marks used together with other code points not (yet) encoded in later revisions in Unicode. The panel will evaluate such cases carefully to ensure freedom of expression for for minority or under-resourced languages in the future.

## 1.3 Countries with significant user communities using Latin script

According to [Wikipedia](https://en.wikipedia.org/wiki/Latin_script) the distribution of the Latin script on the world map is:



Dark green marks countries where the Latin script is the sole main script.

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

Grey marks areas, in which supposedly Latin-script is not used or used only unofficially for second language, however several widely known case which may be eligible for consideration for the LGR run counter to this classification, such as French in Algeria, English in Egypt, or even specialized yet widely used written representations, such as [Arabic chat alphabet](http://www.webcitation.org/6oaqENGQA), or the use of Latin transliteration for Chinese languages, known as P[inyin](https://en.wikipedia.org/wiki/Pinyin).

There are no reliable figures regarding the number of languages using Latin script or the number of readers and writers of the script using community. According to [Worldstandards.eu](http://www.worldstandards.eu/alphabets/), languages using Latin script are spoken by 2.6 billion people (36% of the world population). However such figures just as the visual depictions from Wikipedia are far from exact, since various linguistic factors play into this, including issues of classifications of spoken languages (e.g. dialect vs. language), socio-linguistic factors such as the use of different languages, scripts, and orthographies by the same users for different contexts, as well as issues of language policy, which may or may not recognize or even ban the use of languages, script, and orthographies for political reasons, irrespective of the actual use.

Appendix A is composed using data found on [OMNIGLOT](http://www.omniglot.com/writing/latin.htm) and [ETHNOLOGUE](https://www.ethnologue.com/) sites. However this constitutes only a sample of data. The Generation Panel will consider all languages using Latin script provided they fulfil the criteria of inclusion stipulated by the [Procedure]. In that context, it is particular the criteria of modern wide-spread use which shall serve as crucial factor for inclusion. The Integration Panel [MSR-2] has made use of the EGIDS (Expanded Graded Intergenerational Disruption Scale) as documented in [SIL-Ethnologue] to evaluate modern use, and in [MSR-2] used a cut-off between EGIDS level 4 (Educational) and level 5 (Developing). This however is only seen as an a guideline for inclusion, and where the Panel sees data positively confirming a relevant use of a language of a higher EGIDS rating, the panel shall document such for the final proposal and provide feedback to the Integration Panel within its review of [MSR-2].

# Proposed Initial Composition of the Panel

## 2.1 Panel Chairs and Members

The current working group includes first the members of the panel, then the observers, in alphabetical order:

| **No.** | **Name** | **Position** | **Organization** | **Country** | **Language Expertise** |
| --- | --- | --- | --- | --- | --- |
| 1. | Abdeslam Nasri | Member | ATOS | Algeria | Arabic,French |
| 2. | Ahmed Bakhat Masood | Member | Pakistan Telecom Authority | Pakistan | Urdu, English |
| 3. | Bill Jouris | Member | Inside Products | USA | English, German, Japanese |
| 4. | Dennis Tan Tanaka | Member | Verisign | USA | Spanish, English |
| 5. | Hazem Hezzah | Member | League of Arab States | Egypt | Arabic, German |
| 6. | Jean-Jacques Subrenat | Member | NCUC; Individual Users; NMI/CC; ICG | France | French, English |
| 7. | Mats Dufberg | Member | Internet Foundation In Sweden | Sweden | Swedish, English, some Danish |
| 8. | Meikal Mumin | Member | Institute for African Studies and Egyptology, University of Cologne & “L’Orientale” University of Naples | Germany | German, English, Italian, French, and various African and Middle Eastern languages |
| 9. | Michael Bauland | Member | Knipp  Medien und Kommunikation GmbH | Germany | German, English, Finnish |
| 10. | Mirjana Tasić | Chair | Register of National Internet Domain Names of Serbia (RNIDS) | Serbia | Serbian, Croatian, Bosnian, Montenegrin,  English |
| 11. | Seun Ojedeji | Member | AFRINIC BOD member | Nigeria | English,Yoruba native speaker |
| 12. | Danko Jevtovic | Observer | Register of National Internet Domain Names of Serbia (RNIDS) | Serbia | Serbian, English |
| 13. | Jiankang Yao | Observer | Computer Network Information Center (CNIC, CAS) | China | Mandarin Chinese, Pinyin and English |
| 14. | Matthias Brenzinger | Observer | University of Cape Town | South Africa |  |
| 15. | Tarik Merghani | Observer | AfTLD | Sudan |  |

## 2.2 Relevant expertise

| **N.** | **Name** | **Role** | **Designation** | **Relevant experience** |
| --- | --- | --- | --- | --- |
| 1. | Abdeslam Nasri | ICT Architect/ Arabic Generation Panel | ICT Architect and Project Manager / AtoS | * 2014 to present: Member of the Arabic GP * 2014 to present: Member of the Task Force on Arabic IDN (TF-AIDN) * Expertise in various IT domains like software development, Internet development and multi-tiered architectures, Enterprise architecture. PSPO I and TOGAF certification * Panellist at the Internet Governance Forum |
| 2. | Ahmed Bakhat Masood | Regulator/ DNS/ Arabic Generation Panel/Security | Deputy Director (ICT/Network)/ Pakistan Telecom Authority | * 2013 to present: Member of Task Force on Arabic IDN (TF-AIDN) * 2014- to present: Member of Program Committee Middle East DNS Forum) * 1998 to present: Pakistan Telecom Authority (PTA) * Initiation of different ICT projects for community development like IXP for Pakistan * Coordination for Ipv6 Task Force for Pakistan Network Management, Network Security including DNSSec and Network forensic * Coordination with APNIC, SANOG, ICANN and academia for trainings on modern technologies like IPV6, DNSSec, IRM * Network and Security management * Implementation of ISO 27001 standards in PTA |
| 3. | Bill Jouris | Computer performance measurement and analysis expert | COO of Inside Products | * 2016 to present: technical work on project to identify homographic conflicts in domain names * 2015 to present: work on various IETF RFCs * 2012 to 2015: Director, Computer Measurement Group |
| 4. | Dennis Tan Tanaka | Registry Operator/IDN | Senior Platform Manager | * Member of the IDN Implementation Guidelines Review Team * IDN Product Manager |
| 5. | Hazem Hezzah | Arabic Generation Panel member/ National and regional policy maker | IT Expert for ICT Development / League of Arab States | * 2013-present: Member of the Task Force for Arabic Script IDNs (TF-AIDN) * 2012- present: Member of the Multistakeholder advisory group and preparation team for the Arab Internet Governance Forum. * 2012-present: Participated in preparation, evaluation and contracting for the (.arab) gTLDs, and currently preparing policies for the new gTLD. * 1991-2011: Performed various IT related roles as support, consultant and technical project manager. * Languages: English, German, use of Latin script for Arabic chat language |
| 6. | Jean-Jacques Subrenat | Policy Expert | Président, IndividualUsers.org (European Individual Users’ Association) | * Member of the NTIA IANA Functions’ Stewardship Transition Coordination Group (ICG) * Member of the NETMundial Coordination Council * President of the Steering Committee, IndividualUsers.org (elected in October 2015) Member of the ICANN Board of Directors 2007-10 during which: * Member of President’s Strategy Committee (where he was a co-author of the “Implementation Plan for Improving Institutional Confidence”) * Structural Improvements Committee; Public Participation Committee (as its first Chair) * Member of Board Working Groups: ALAC Review, Board Review, ccNSO Review (as its Chair) |
| 7. | Mats Dufberg | IDN/DNS/  Linguist | Internet Foundation In Sweden | * DNS specialist * New GTLD Pre-Delegation testing * BA Phonetics/Linguistics |
| 8. | Meikal Mumin | Linguist | Institute for African Studies and Egyptology, University of Cologne & “L’Orientale” University of Naples | * Member of Arabic Generation Panel * Member of Task Force on Arabic Script IDNs (TF-AIDN) * PhD candidate and researcher with an expertise in Writing Systems and orthographies, in particular such used in Africa and the Middle East. * User of several Modern European Languages |
| 9. | Michael Bauland | DNS/Registry/ Registrar / IDN | Knipp Medien und Kommunikation GmbH | * Development of IDN table for <بازار‎> (.bazaar) * 2007-present: Senior Software Engineer at Knipp * 2003-2007: Senior Research Assistant at Leibniz University of Hanover |
| 10. | Mirjana Tasić | Registry / DNS/Unicode Expert / IDN | Executive Advisor, RNIDS (Register of National Internet Domain Names of Serbia) | * 08/2012–12/2012 ICANN IDN variant TLD Program: Project (P2.1) – Procedure to Develop and Maintain the Label Generation Rules for the DNS Root Zone in Respect of IDNA Labels -ICANN volunteer * Introduction and implementation of IDN ccTLD Fast Track Process for ccTLD <срб><xn—90a3ac>: string evaluation, domain delegation, sunrise and open registration. * 07/2006–03/2009 Acting Director of RNIDS (volunteer work). Preparation and implementation of .rs landrush procedures; organization and implementation of the transition process from .yu to .rs domain. * 04/2006–07/2006 Founder of RNIDS (volunteer work). * 04/1994–09/2008 YU TLD (YU Top Level Domain) Administrator (volunteer work). Managed operation of .yu DNS; Maintained database of .yu domains. * 1992–1994 Chairwoman, Technical Committee, Academic Network of Yugoslavia. Actively participated in the introduction of internet in Serbia. (volunteer work) * 1991–10/2010 Administrator of Class B IP address (147.91) assigned to the University of Belgrade, Serbia. (volunteer work) |
| 11. | Seun Ojedeji | DNS/ Policy / Native speaker | Chief Network Engineer at Federal University of Oye-Ekiti | * AFRINIC Member Board of Directors * Open Source Foundation for Nigeria Vice President * FOSSFA Council Chair * AFRNIC Policy Development Working Group co-chair * Principal Network Engineer * System analyst/Network engineer |
| 12. | Danko Jevtovic | Observer |  |  |
| 13. | Jiankang Yao | Observer |  |  |
| 14. | Matthias Brenzinger | Observer |  |  |
| 15. | Tarik Merghani | Observer |  |  |

## 2.3 Panel Diversity

The panel currently includes members from several areas of the script using community. However as the Latin script is used by hundreds of languages, it is not possible to have direct representation from community members or speakers from all of them. At the same time, there are numerous languages, for which resources are scarce or difficult to access. Where the panel finds it lacks sufficient expertise and sees an actual need, it will work together with ICANN and request support in obtaining help from external experts, who are not formally members of the panel, to act as advisors to the panel following the [Procedure].

Those individuals listed without short CVs and currently marked as observers have at some point expressed their willingness to support the work of the panel, but have not reacted to further communiques during the process leading to the seating of the proposed panel and the submission of the present proposal. In the interest of openness and because of the panel’s wide remit, the members of the suggested panel have decided not to remove those individuals for the moment, and the panel expressly remains open to new members or observers becoming full members throughout its work.

| **Diversity** | **Panel members** |
| --- | --- |
| Community representatives | * Jean-Jacques Subrenat * Mirjana Tasić |
| Linguistic expert | * Mats Dufberg * Meikal Mumin |
| Technical community (general and DNS) | * Abdeslam Nasri * Ahmed Bakhat Masood * Bill Jouris * Dennis Tan Tanaka * Mats Dufberg * Meikal Mumin * Michael Bauland * Mirjana Tasić * Seun Ojedeji |
| Registry/Registrar expert | * Dennis Tan Tanaka * Michael Bauland * Mirjana Tasić |
| Policy expert | * Ahmed Bakhat Masood * Jean-Jacques Subrenat * Mirjana Tasic |
| Local language computing using Unicode and specifically IDNs | * Ahmed Bakhat Masood * Dennis Tan Tanaka * Hazem Hezzah * Mats Dufberg * Meikal Mumin * Michael Bauland * Mirjana Tasić * Seun Ojedeji |

## 2.4 Relationship with Past Work or Working Groups

ICANN’s Variant Issues Project Study Group for the Latin Script produced Considerations in the use of the Latin script in variant internationalized top-level domains in 2011.

The Generation Panel intends to take into consideration and investigate recommendations from other relevant Panels such as the Integration Panel as well as from other Generation Panels which have contributed prior relevant work, particularly when making choices about the repertoires and variant relationships within, such as the “feasibility and risks of supporting the sharp s in the LGR” [MSR-2: 19] and, if it should consider the inclusion of this code point in the LGR, to investigate the case for or against making it a blocked variant of 'ss', or the view of Armenian GP regarding relevant cross-script variants.

# Work Plan

The role of the Latin Generation Panel is to establish the repertoire and Label Generation Rules for top level internationalized domain names in Latin script.

## 3.1 Suggested Timeline with Significant Milestones

The Generation Panel intends to divide the work on the LGR for the Root Zone into five stages:

1. Development of the Code Point Repertoire
2. Development of the Code Point Variants
3. Development of the Whole Label Evaluation Rules
4. Preparation of the Latin script LGR for public comment
5. Finalization of the LGR for Latin Script and submission to ICANN

At all stages there will be consultation with the Integration Panel and the public via periodic public comment phases.

|  | **Activity** | **Completion** |
| --- | --- | --- |
| **1.** | **Developing Code Point Repertoire** |  |
|  | Determine criteria for including code points | 4 weeks |
|  | Analyze each code point in Latin script in MSR-2 to determine its inclusion  Analyze languages using Latin script to determine if all relevant letters are included in MSR-2 | 18 weeks |
|  | Finalize included and excluded code points | 2 weeks |
|  | Release for public comment by Latin GP | 4 weeks |
|  |  |  |
| **2.** | **Developing code point variants** |  |
|  | Determine criteria for within-script variants | 4 weeks |
|  | Determine criteria for cross-script variants, including identification of relevant scripts | 2 weeks |
|  | Analyze each code point to determine within-script variant sets | 8 weeks |
|  | Analyze each code point to determine cross-script variant sets | 8 weeks (in parallel with within-script variants above) |
|  | Review the impact of variant analysis on current delegations of TLDs | 4 weeks |
|  | Finalize variant sets, reviewing the symmetry, transitivity, security and stability of the system | 4 weeks |
|  | Coordinate with other relevant GPs for finalization of cross-script variants | 4 weeks (in parallel with the Public Comment phase below) |
|  | Release for public comment by Latin GP | 4 weeks |
|  |  |  |
| **3.** | **Discuss WLE rules needed for Latin script LGR** |  |
|  | Determine relevant WLE rules | 4 weeks |
|  |  |  |
| **4.** | **Prepare Latin script LGR proposal for public comment** |  |
|  | Finalize the documentation for Latin script LGR | 4 weeks |
|  | Finalize the XML formulation of the proposal | 4 weeks (in parallel to documentation) |
|  | Collect labels to test the proposed LGR | 2 weeks (in parallel to the the documentation) |
|  |  |  |
| **5.** | **Finalize Latin LGR proposal** |  |
|  | Release for formal public comment by ICANN | 6 weeks |
|  | Finalize Latin LGR proposal based on feedback from the public comment phase | 4 weeks |

Note that this schedule will be updated in a timely manner according to the increasing/decreasing number of action items, and the coordination situation.

The panel will hold fortnightly the conference calls with agenda and materials ready for discussion and all working materials will be shared online.

It is expected that the panel will have few face-to-face meetings, which will take place probably during phases 2 and 3.

## 3.2 Sources for funding travel and logistics

Although the members of the Latin panel will be volunteers and provide their time and

expertise on a purely voluntary basis, issues of logistics such as travel and stay necessitated by

members in a face-to-face meeting as well as support for conference calls, and postings on the website will require support.

## 3.3 Need for ICANN provided advisors

The panel might need ICANN support for stage 1 **Developing Code Point Repertoire.** Support might be necessary in analyzing the code point repertoire for all languages of the script using community, a sample of which is presented in Appendix A.

Some coordination might also be needed with Greek and Cyrillic GP.

# References

1. Frakes, J., et al., “Considerations in the use of the Latin script in variant internationalized top-level domains: Final report of the ICANN VIP Study Group for the Latin script”. Los Angeles, Calif.: ICANN, October 2011). <http://archive.icann.org/en/topics/new-gtlds/latin-vip-issues-report-07oct11-en.pdf>
2. Blanchet, M., et al. "Guidelines for Developing Script‐Specific Label Generation Rules for Integration into the Root Zone LGR". Los Angeles, Calif.: ICANN, April 2015. <https://community.icann.org/download/attachments/43989034/Guidelines%20for%20LGR.pdf>
3. "Considerations for Designing a Label Generation Ruleset for the Root Zone". Los Angeles, Calif.: ICANN, April 2015. <https://community.icann.org/download/attachments/43989034/Considerations%20for%20LGR.pdf>
4. "Requirements for LGR Proposals". Los Angeles, Calif.: ICANN, April 2015. <https://community.icann.org/download/attachments/43989034/Requirements%20for%20LGR%20Proposals.pdf>
5. Integration Panel “Root Zone Label Generation Rules – LGR-1 Overview and Summary”. Los Angeles, Calif.: ICANN, February 2016.
6. Common Locale Data Repository. [www.unicode.org/cldr/charts/28/summary/root.html](http://www.unicode.org/cldr/charts/28/summary/root.html)
7. [www.ethnologue.com](http://www.ethnologue.com)
8. [www.omniglot.com](http://www.omniglot.com)
9. [www.scriptsource.org](http://www.scriptsource.org)
10. <https://en.wikipedia.org/wiki/History_of_the_Latin_alphabet>
11. <https://en.wikipedia.org/wiki/Latin_script>
12. Maximal Starting Repertoire [MSR-2]. <https://www.icann.org/resources/pages/reports-2013-04-03-en>
13. Katz, Leonard & Ram Frost. 1992. The Reading Process is Different for Different Orthographies: The Orthographic Depth Hypothesis. *Haskins Laboratories Status Report on Speech Research* 111/112. 147–160.
14. Klensin, J., “Internationalized Domain Names in Applications (IDNA): Definitions and Document Framework” = RFC 5890 (2010). <http://tools.ietf.org/html/rfc5890>
15. Fältström, P., ed., “The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)” = RFC 5892 (2010). <http://tools.ietf.org/html/rfc5892>
16. Hoffman, P., et al., “Terminology Used in Internationalization in the IETF” (2011). = RFC 6365 <http://tools.ietf.org/html/rfc6365>
17. [Procedure] Sullivan, A., et al., “Procedure to develop and maintain Label Generation Rules for the Root Zone in respect of IDNA labels” (Marina del Rey, California: ICANN, March 2013). [https://www.icann.org/en/system/files/files/lgr-procedure-20mar13-en.pdf.](https://www.icann.org/en/system/files/files/lgr-procedure-20mar13-en.pdf)
18. Bendor-Samuel, John. 1996. African languages. In Peter T. Daniels & William Bright (eds.), *The World's Writing Systems*, 689–691. New York, Oxford: Oxford University Press.
19. Hartell, Rhonda L. (ed.). 1993. *Alphabets de Langues Africaines*. Dakar: UNESCO, Bureau Régional de Dakar; Societe Internationale de Linguistique.
20. IDNA 2008. See RFCs 5890, 5891, 5892, 5893, 5894 and 5895. <https://tools.ietf.org/html/rfc5890> <https://tools.ietf.org/html/rfc5891> <https://tools.ietf.org/html/rfc5892> <https://tools.ietf.org/html/rfc5893> <https://tools.ietf.org/html/rfc5894> <https://tools.ietf.org/html/rfc5895>
21. ISO 15924 “Codes for the representation of names of scripts”. <http://unicode.org/iso15924/iso15924-codes.html>
22. [RFC1034] Mockapetris, P., "Domain names - concepts and facilities", STD 13, RFC 1034, November 1987.
23. [RFC1035] Mockapetris, P., "Domain names - implementation and specification", STD 13, RFC 1035, November 1987.
24. [Armenian Proposal] Armenian Generation Panel (GP). Proposal for an Armenian Script Root Zone Label Generation Rules”, 22 Jul 2015. <https://www.icann.org/public-comments/proposal-armenian-lgr-2015-07-22-en>
25. [SIL-Ethnologue] Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2014. Ethnologue: Languages of the World, Seventeenth edition. Dallas, Texas: SIL International. Online version available as http:[//www.ethnologue.com.](http://www.ethnologue.com.)
26. [Unicode Standard Annex #15]: <http://unicode.org/reports/tr15/>
27. [Punycode]: Costello, A., “Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA)”, RFC 3492, March 2003, <https://tools.ietf.org/html/rfc3492>

1. *Script* is used here to indicate the whole writing system including basic letters, ligatures and diacritics. See also RFC 6365 and ISO 15924. [↑](#footnote-ref-1)
2. However, several orthographies on the basis of different scripts are frequently used simultaneously, both historically and contemporarily. [↑](#footnote-ref-2)
3. In writing and typography, a ligature occurs where two or more graphemes or letters are joined as a single glyph. [↑](#footnote-ref-3)
4. Due to the variation of language names, ISO 639-3:2007 codes are placed in brackets after such language names in the present document to aid in identifying the correct idiom. [↑](#footnote-ref-4)
5. The exact representation of linguistic features however is different for every orthography: Some languages, such as Esperanto, use it more phonemically, while other languages, such as English, use it so that other aspects, such as etymology, are represented too. For example, the spelling of ‘night’ connects it with German ‘Nacht’, although 'gh' is no longer pronounced. Therefore, the degree to which phonemes are represented can vary from orthography to orthography, and this characteristic has been described by the term orthographic depth in linguistic literature (cf. Katz & Frost 1992). [↑](#footnote-ref-5)
6. The Combining stroke is not encoded as a combining mark in Unicode. [↑](#footnote-ref-6)
7. Few diacritics have not been encoded separately in Unicode but only as complex glyphs such as the combining stroke of Navajo given above. [↑](#footnote-ref-7)
8. The Greek, Arabic and Hebrew scripts developed from the Phoenician alphabet, but the relationship is so distant that there is little visual similarity among most related letters among them. [↑](#footnote-ref-8)