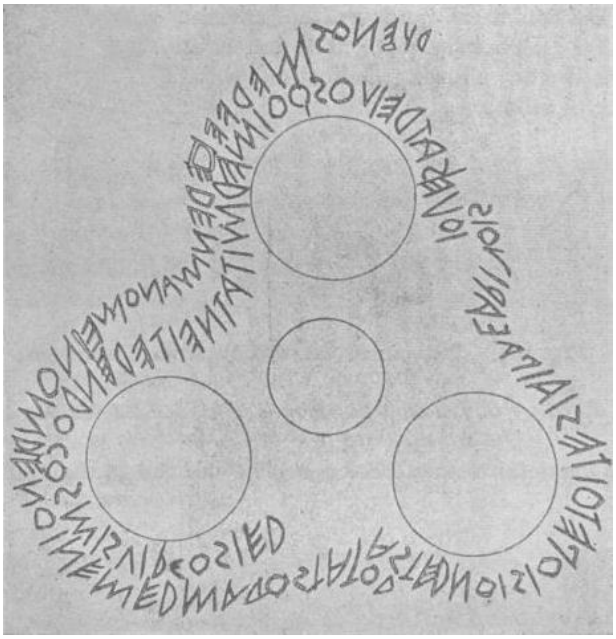
Proposal for Generation Panel for Latin Script Label Generation Ruleset for the Root Zone. Ed. C. Dillon. Version 3 (14 Jan. 2015)

1. General Information

The Latin script 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.

The Latin script originated in Italy in the 7th Century BC. The original letters were: A, B, C, D, E, F, Z, H, I, K, L, M, N, O, P, Q, R, S, T, V and X. There were only upper case letters with serifs.

G developed from C and J from I. V and U split and a ligature of VV became W. Languages added new letters, for example þ (thorn) for Scandinavian languages, borrowed from the runic alphabet. Letters were often combined to form ligatures, for example æ (from a and e, in e.g. Danish and Norwegian) or ß (from Gothic s and z, in German). The current basic set is: 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.



The Latin script is alphabetic – there are letters for both consonants and vowels. Some languages, such as Esperanto, use it phonemically, so that sounds are represented in a systematic way; other languages, such as English, use it so that other aspects, such as etymology, are represented.

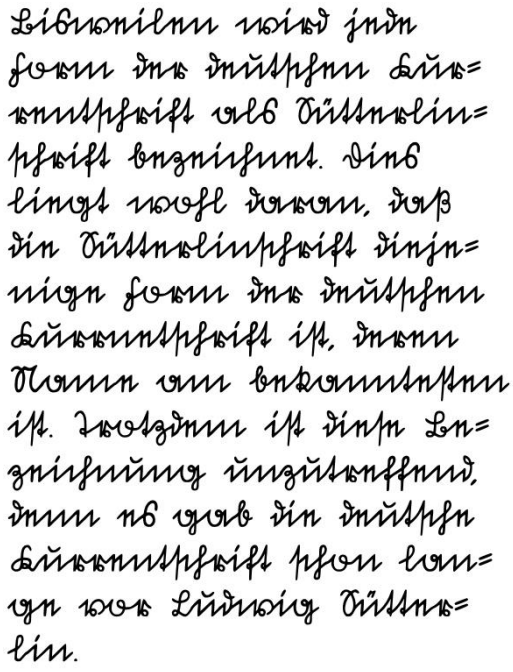
The Latin script exists in both upper and lower case forms. There may be little visual similarity between a letter's upper and lower case form, for example, A and a.

It is almost always written left-to-right.

The Duenos Inscription, 6th Century B.C.

Letter shapes may be considerably different depending on the language and whether the script is handwritten or printed.

There are many different writing styles. Until the 1940s, for example, German was commonly written in Gothic (or blackletter) script ("Fraktur"). Sütterlin was a common form.



sample of Fraktur by -donald-

Normally letters appear separately when printed and joined together when written by hand. However, some printed fonts join the letters together and many people have individual preferences to write at least some letters separately in their handwriting.

Spaces are almost always used to separate words. The hyphen (-) is used in many languages to separate elements that belong together in some way, for example, parts of a compound noun or to indicate that a word has been truncated, for example, at the end of a line.

Diacritics and marks also came to be used to modify letters in many languages. These may appear anywhere around, most commonly above (é), below (ç), or through (ø) a letter. Several diacritics may attach to the same letter; Vietnamese ợ, for example, has a hook on the right and a dot below.

Some languages consider letter + diacritic as one letter. Norwegian (both Bokmål and Nynorsk varieties), for example, lists these three letters at the end of its alphabet: Æ, Ø and Å.

Diacritics may perform different roles depending on the language:

- For example, in French the acute accent over e (é) is used to indicate a closed e sound, for example café.
- In Spanish, however, the same diacritic is used to indicate cases where the stress does not fall on the penultimate syllable, for example, dieciséis 'sixteen'.
- In Vietnamese, the same diacritic would indicate a high rising tone.

As represented in Unicode, the Latin script has some identical glyphs, for example, 0259 ə (schwa) and 01DD ǝ (turned e). The following glyphs belong to both the Latin and Cyrillic scripts: A, B, E, S, I, J, K, M, O, C, T, У, and X (this is a non-exhaustive list of the Cyrillic variants). In the case of H and P it should be noted that they represent N and R sounds in Cyrillic.

A letter with two diacritics, for example, ḉ may be typically represented in several ways in Unicode – as a precomposed form (U+1E09), or as the letter and the first diacritic with the second added (U+0107 ć + U+0327 ◌̧ COMBINING CEDILLA), or with the letter and the second diacritic with the second diacritic added (U+00E7 ç + 0301 ◌́ COMBINING ACUTE ACCENT).

It is possible that combining marks may be required for some languages in widespread modern use.

Code points must be in contemporary use, and must not be punctuation, or solely for historical, religious text or other specialist use.

To determine whether a code point is in a language in modern use, websites such as Ethnologue including EGIDS (Expanded Graded Intergenerational Disruption Scale), Omniglot and Unicode (especially the Common Locale Data Repository) and Wikipedia were used. Other major criteria include the number of speakers and whether there exists a modern literature or newspapers in the language.

1.1 Target Script for the Proposed Generation Panel

Latin script has the following specifications:

ISO 15924 code: Latn

ISO 15924 no.: 215

English Name: Latin

Note that the Gaelic and Fraktur variants of Latin have their own ISO 15924 codes and numbers (Latg 216 and Latf 217 respectively), and so do not fall within the remit of the LGP.

The complete set of characters in the Latin script fall in the following Unicode ranges:

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
Latin Extended-C	U+2C60 – U+2C7F
IPA Extensions	U+0250 – U+02AF
Combining Diacritical Marks	U+0300 – U+036F
Latin Extended-D	U+A720 – U+A7FF
Combining Diacritical Marks Supplement	U+1DC0 – U+1DFF
Latin Extended Additional	U+1E00 – U+1EFF
Latin Ligatures	U+FB00 – U+FB0F
Full-width Latin Letters	U+FF00 – U+FF5E

MSR2 excluded the following ranges:

- Latin Extended-D; technical use (phonetic)/obsolete/punctuation
- Latin Ligatures; ##
- Full-width Latin letters; ##

It is possible that there may not be precomposed forms in Unicode 8.0 for all letters in languages in modern use or even letters that cannot be represented.

Certain characters are only used for historical purposes. For example, some consonants in Irish Gaelic were formerly written with a dot above them, e.g., ḃ, ċ, ḋ, ḟ, ġ, ṁ, ṗ, ṡ and ṫ; now they are written: bh, ch, etc.

The Latin script is often used to Romanize other languages. For example, in the Hepburn Romanization of Japanese, 東京 would be written as Tōkyō. Romanization may require the use of unusual diacritics, for example, a dot under a consonant (ḍ) may represent a retroflex sound, as in Indian and Scandinavian languages. Many unofficial Romanizations also exist such as Arabic chat:
ana raye7 el gam3a el sa3a 3 el 3asr.

1.2 Foundation documents and RFCs

Terminology Used in Internationalization in the IETF (RFC 6365) is used for definitions.

The normative statement of the protocol-valid code points is given in RFC 5892 with a corresponding reference table in the IANA Protocol Registry.

Code points in IDNs in Latin script must be PVALID in the IDNA 2008 protocol and CONTEXT O/J.

1.3 Principal languages using the script

Major world languages using the Latin script include:

- Europe: Many Romance, Germanic and Slavonic, and some other languages including Spanish, French, Italian, Portuguese, English, German, Dutch, Swedish, Danish, Norwegian, Polish, Czech, Croatian, Finnish and Hungarian.
- America: Many European languages plus indigenous languages including ##
- Eskimo-Aleut: ##
- Africa: Many European languages plus indigenous languages including Swahili, Hausa and Yoruba.
- Central Asia: Azeri, Turkish, Turkmen, Uzbek, etc.
- Australasia and South-East Asia: Many European languages plus Pitjantjatjara, Maori, Indonesian, Bahasa Malaysia, Tagalog, Vietnamese, etc.

See Appendix A for a longer but probably non-exhaustive list.

Europe

- The Latin script is the writing system in widest use in Europe. Cyrillic is used by several countries, for example Bulgaria and Serbia and the Greek alphabet is used by Greece.

- Many languages have modified letters by adding diacritics, for example, ą in Polish or created digraphs, for example, œ in French or new letters, for example þ (thorn) in Icelandic.

America

##

Eskimo-Aleut

##

Africa

- Today, the Latin script is the writing system in widest use in Africa.
- It is estimated that over 500 out of the 2000 languages spoken in Africa today have orthographies (Bendor-Samuel 1996: p.689), with the vast majority being Latin script-based.
- The Latin script has been significantly extended or modified to represent African languages:
    - Frequently, supra-segmental features such as tone were encoded using super-and subscripted graph(eme)s, such as accent marks.
    - Next to entirely new letters, di-, tri- and quadrigraphs, for example, are often-much used to represent single phonological units.
- A number of code-points are already excluded by the "letter principle" in the MSR, as well as IDNA 2008.

Central Asia ##

- The languages of the majority of the inhabitants are Turkic: Azeri, Tatar, Turkish, Turkmen, Uzbek, etc.
- Some languages in the area are sometimes and other exclusively written in the Cyrillic or Arabic scripts.
- Some diacritics are used, for example, ü and ş in Azeri, Turkish and Turkmen, and some additional letters are used, for example, ə (schwa) in Azeri.

South East Asia and Australasia

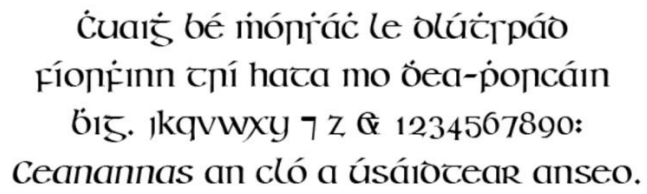
This area contains Polynesian, Australian, Austronesian and Papuan languages. Major Polynesian languages include Hawaiian, Maori, Samoan, Tahitian and Tongan. Long vowels may be indicated by macrons, for example, ō.

- There are fewer than 150 Australian languages in modern use. Some use digraphs, and some diacritics, for example n̲ in Pitjantjatjara.
- There are over 1,000 Austronesian languages, including Bahasa Malaysia, Indonesian, Formosan languages and Tagalog. Most Austronesian languages now use the Latin script, but there is some use of the Arabic script, for example, Jawi.
- Some Austronesian languages are spoken in New Guinea. Most of the over 1,000 languages spoken there are Papuan languages with Latin-based writing systems.

1.4 Related Scripts

As mentioned above, the Latin and Cyrillic scripts developed from the Greek script and share several letters. 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. The Armenian script may be modelled on the Greek script and a small number of letters are shared.

The Fraktur and Irish Gaelic writing styles of the Latin script are so different that Unicode considers them different scripts.



sample of Irish Gaelic by Arthur Baker

2. Proposed Initial Composition of the Panel

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

2.1 Panel Chairs and Members (with Expertise)

The current working group includes the following members in alphabetical order:

No.	Name	Position	Organization	Country	Language Expertise
1	Tunde Adegbola	Member	African Languages Technology Initiative	Nigeria	
2	Sarat Assirou	Member	Institute COADY	Ivory Coast	Dioula
3	Dwayne Bailey	Member	Translate.org.za	South Africa	Afrikaans, Northern Sotho, Venda, Tswana and Southern Sotho
4	Ahmed Bakht Masood	Member	Pakistan Telecom Authority	Pakistan	Urdu, English
5	Matthias Brenzliger	Member	University of Cape Town	South Africa	
6	Chris Dillon	Co-Chair	University College London	UK	English, German, Spanish
7	Tarkan Doruk	Member	Sanofi	UAE	Turkish

8	Yashar Hajiyev	Member	Information Policy Analytical Center	Azerbaijan	Azerbaijani, English
9	Hazem Hezzah	Member	League of Arab States	Egypt	Arabic, German
10	Paul Hoffman	Member	ICANN	US	English
11	Tarik Merghani	Member	AfTLD	Sudan	
12	Meikal Mumin	Member	University of Cologne	Germany	German, English, use of Latin script for African languages
13	Danko Jevtovic	Member	Fondacija	Serbia	Serbian, English
14	Ngo Thanh Nhan	Member	New York University	US	Vietnamese
15	Daniel Omondi	Member	Internet Society	Kenya	
16	Oscar Gabriel Ledesma Piñeiro	Member	Alfa-REDI	Argentina	Spanish, English
17	Gideon Kiprono Rop	Member	DotConnectAfrica	Kenya	
18	Jean-Jacques Subrenat	Member	NCUC; Individual Users; NMI/CC; ICG	France	French, English
19	Mirjana Tasić	Member	National Internet Domain Names of Serbia (RNIDS)	Serbia	Serbian, English
20	Aysegul Tekce	Member	ICANN	Turkey	Turkish
21	Eric Brunner-Williams	Co-Chair	CORE	US	English
22	Bonface Witaba	Member	Global Knowledge Partnership Foundation	Kenya	Swahili
23	Jiankang Yao	Member	Computer Network Information Center (CNIC, CAS)	China	Mandarin Chinese, Pinyin and English

Relevant expertise## — summaries of experience from CVs in bullet point-form.

2.2 Panel Diversity

As the Latin script is used by several hundred languages (see the appendix), it is not possible to have representation from experts for all of them. The approach taken, therefore, is to have experts covering areas of languages, for example, African languages using the Latin script.

National and regional policy makers

Some members of the panel are well versed in ICANN policy, others in national and regional policy.

Technical community (general and DNS)

One member of the panel has a strong technical background (including RFC-writing).

Security and law enforcement

##

Academia (technical and linguistic)

The panel has good coverage of European languages (Romance, Germanic and Slavonic), some coverage of North American indigenous languages, some coverage of African languages, but only weak coverage of South East Asian and especially Central Asian languages and again weak coverage of Australasian languages.

Community-based organizations

Several members of the panel work for community organizations.

Local language computing using Unicode and specifically IDNs

Several of the linguists have a good knowledge of local language computing, Unicode, IDNA and ICANN's Variant Issues Project.

2.3 Relationship with Past Work or Working Groups

Until the advent of IDNs in 2003, the "LDH set" – Latin letters "a" to "z" in both upper and lower case, the digits "0" to "9" and the hyphen was used for the registration of names in the DNS.

IDNA (Internationalized Domain Names in Applications) is the protocol used for implementing IDNs. The latest version is 2008, but changes from the 2003 version are likely to break the Longevity Principle in the *Procedure to develop and maintain Labe Generation Rules for the Root Zone in respect of IDNA labels*.

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.

3. Work Plan

3.1 Suggested Timeline with Significant Milestones

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

1. Finalization of Code Points
2. Finalization of Variants (if any)
3. Finalization of Whole Label Rules
4. Finalization of LGR Documents for Latin Script and Submission to ICANN

At all stages there will be consultation with the Integration Panel, the Generation Panels of Related Scripts, and the public via periodic public comments.

1. Finalization of Code Points

This stage involves the listing of code points from the parts of Unicode listed in section 1.1 above. Rows in the list will be coloured red until they are attested as in languages in modern use, at which point the colouring will be changed to black. This situation will be represented in an XML file. For the non-exhaustive list of languages using the Latin script that is to be used, see the appendix.

2. Finalization of Variants (if any)

The LGP will decide if it is necessary to declare in-script and/or cross-script variants. In the event of a declaration of either sort of variant, an exhaustive list will be made. This situation will be represented in an XML file.

3. Finalization of Whole Label Rules

The LGP will check that no problems are caused by any default WLE and then list any Latin script-specific WLEs, if, for example, some code point may only occur in certain positions in a label, or may only occur together with certain other code points or ranges of code points. This situation will be represented in an XML file.

4. Finalization of LGR Documents for Latin Script and Submission to ICANN

The proposal document and XML files will be completed, taking into account public comments and the work of the Generation Panels of related scripts (at least Greek and Cyrillic). It is possible that a delay may be necessary at this stage.

3.2 Proposed schedules of meetings and teleconferences

The schedule below roughly presumes the Arabic Generation Panel's schedule. The AGP's experience is likely to speed up the LGP's work. The Latin script, however, is used by a larger number of languages and consists of a larger number of code points; both factors which will slow down the LGP's work. The schedule presumes about four months on work with variants. If they are not declared, this may decrease to as little as one month. It may be necessary to appoint advisors to fill gaps in the panel's experience. The panel is composed largely of volunteers and not all of them will have time at all stages of the work.

Task name	**By**	**Status**
Develop call for participation	Tue 06-23-15	Done
Publicly release call for participation	Fri 07-24-15	Done
Meeting	Tue 9-22-15	Done
Face-to-face meeting (Dublin)	Sun 10-18-15	Done
Meeting on character set	Tue 11-10-15	Done
Invitation to experts to ensure diversity	Fri 11-20-15	In progress
Meeting on character set	Tue 11-24-15	Done
Meeting on character set	Tue 12-08-15	Done
Meeting on panel-formation proposal	Tue 01-05-16	Done
Meeting on panel-formation proposal	Tue 01-**26**-16	
Meeting on panel-formation proposal	Tue 02-**09**-16	
Meeting on finalization of membership	Tue 02-**23**-16	
Proposal finalization	Fri 02-**25**-16	
Application to ICANN for formation of LGP	Fri 02-26-16	
Meeting	Tue 03-01-16	
Face-to-face meeting (Marrakech)	Sun 03-06-16	
Meeting on character set	Tue 03-22-16	
Meeting on character set	Tue 04-12-16	
Meeting on general principles for inclusion	Tue 04-26-16	
Meeting on general principles for exclusion	Tue 05-10-16	
Meeting on general principles for deferral	Tue 05-24-16	
Release of character set for public comment	Tue 06-07-16	
Meeting	Tue 06-21-16	
Face-to-face meeting (Panama)	Mon 06-27-16	
Meeting on finalization of character set	Tue 07-12-16	

Meeting: Discussion on variants	Tue 07-26-16	
Meeting: Are in-script variants needed?	Tue 08-09-16	
Meeting: Are cross-script variants needed?	Tue 08-30-16	
Meeting	Tue 09-13-16	
Meeting on finalization of variants	Tue 09-27-16	
Meeting: Release of variants for public comment	Tue 10-25-16	
Face-to-face meeting (Puerto Rico)	Sun 10-29-16	
Incorporation of comments from public and IG	Tue 11-29-16	
Finalization of variants	Tue 12-13-16	
Discussion of Whole Label Rules	Tue 01-10-17	
Documenting Whole Label Rules	Tue 01-24-17	
Meeting	Tue 02-07-17	
Meeting on finalization of Whole Label Rules	Tue 02-21-17	
Release of Whole Label Rules for public comment	Tue 03-07-17	
Face-to-face meeting (Europe)	Sun 03-12-17	
Incorporation of comments from public and IG	Tues 03-21-17	
Finalize document	Tues 04-04-17	
Meeting	Tues 04-18-17	
Finalize LGR XML structure	Tues 05-02-17	
Final edits	Tues 05-16-17	
Submission to ICANN	Tues 05-30-17	

4. References

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

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

"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

"Requirements for LGR Proposals". Los Angeles, Calif.: ICANN, April 2015. https://community.icann.org/download/attachments/43989034/Requirements%20for%20LGR%20Proposals.pdf

Common Locale Data Repository. www.unicode.org/cldr/charts/28/summary/root.html

www.ethnologue.com

www.omniglot.com

https://en.wikipedia.org/wiki/History_of_the_Latin_alphabet

https://en.wikipedia.org/wiki/Latin_script

Maximal Starting Repertoire (MSR2). https://www.icann.org/resources/pages/reports-2013-04-03-en

https://en.wikipedia.org/wiki/Sütterlin

https://en.wikipedia.org/wiki/Gaelic_type

Klensin, J., "Internationalized Domain Names in Applications (IDNA): Definitions and Document Framework" = RFC 5890 (2010). http://tools.ietf.org/html/rfc5890

Fältström, P., ed., "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)" = RFC 5892 (2010). http://tools.ietf.org/html/rfc5892

Hoffman, P., et al., "Terminology Used in Internationalization in the IETF" (2011). = RFC 6365 http://tools.ietf.org/html/rfc6365

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

Bendor Samuel, J., "African languages" (1996 p.689-691). Oxford University Press

Hartell, R.L., ed., "Alphabet de langues africaines". UNESCO - Bureau Regional de Dakar, 1993