Proposal for the Thai Script Root Zone LGR

LGR Version 2

Date: 2016-12-08

Document version: 6.3

Authors: The Generation Panel for the Thai Script LGR

Contents

1 General Information/ Overview/ Abstract

The purpose of this document aims to give an overarching view of the label generation rules for the Thai Script including rationale behind the design decisions taken. This includes a discussion of the relevant features of the script, the communities and languages using it, as well as the process and methodology used and information of the contributors. The formal specification of the LGR can be found in the accompanying XML document:

- Proposed-LGR-ThaiScript-20161208.xml

Labels for testing can be found in the accompanying text document:

- Labels-ThaiScript-20161208.txt

2 Script for which the LGR is proposed

ISO 15924 Code: Thai

ISO 15924 Key Number: 352

ISO 15924 English Name: Thai

Native name of the script: ไทย

Thai Unicode range: U+0E00 – U+0E7F

This Thai Script Root Zone LGR is based on Maximal Starting Repertoire version: MSR-2

3 Background on Script and Principal Languages Using It

Thai is the official language of Thailand. The Thai script system has been used for Thai, Pali, and Sanskrit languages in Buddhist texts all over the country. The standard Thai is used in all schools in Thailand, and most dialects of Thai use the same script.

In addition, there is also several other languages in Thailand, Laos, Myanmar and China written in the Thai script, however, with moderate to small number of population. Therefore, the panel mainly considers working on the Thai script for the languages that are still actively used and has been reported as having high-usage population (not less than 1 million) as listed in Table 1.

Language	ISO 639-3 Code	Locations	Population (in all countries)	EGIDS Score	Language Name in the Thai Script
Thai	tha	Thailand (official language of Thailand)	60,489,750 as L1: 20,489,750; as L2: 40,000,000 (2000)	1	ภาษาไทย
Northeastern Thai	tts	Widespread in Northeast Thailand	15,000,000 (1983 SIL)	6a	ภาษาอิสาน

Language	ISO 639-3 Code	Locations	Population (in all countries)	EGIDS Score	Language Name in the Thai Script
Northern Thai	nod	Northern region of Thailand	6,000,000 (1983 SIL)	5	ภาษาคำเมือง
Southern Thai	sou	Southern region of Thailand	4,500,000 (2006 Mahidol University)	5	ภาษาปักษ์ใต้
Northern Khmer	kxm	Northeastern and Eastern regions of Thailand along the border with Cambodia	1,400,000 (2006 Mahidol University)	5	ภาษาเขมรเหนือ
Pattani Malay	mfa	Southern region of Thailand near the border with Malaysia	1,000,000 (2006 Mahidol University)	5	ภาษายาวี

Table 1: Selected Languages written in the Thai Script

3.1 Thai

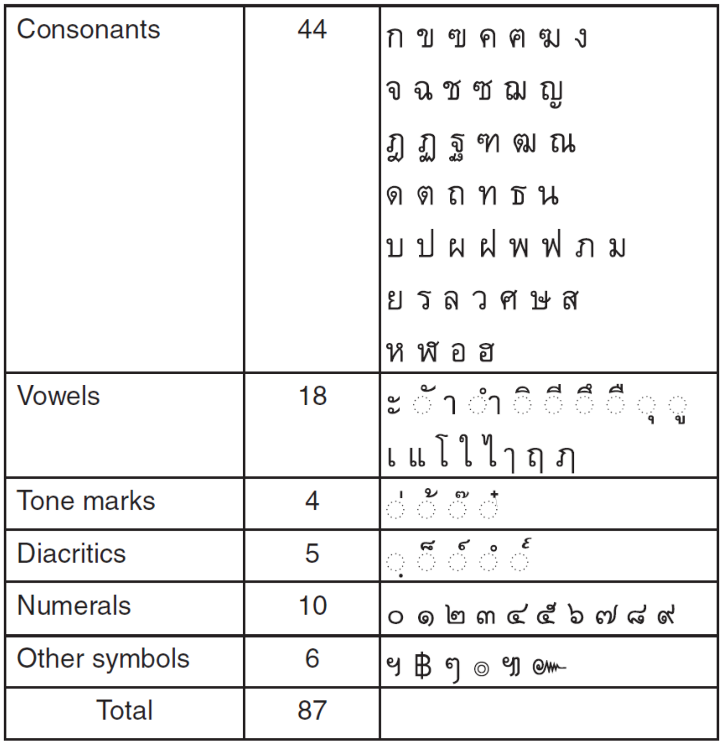
Thai is the language of 65 million people, and has a number of regional dialects, such as Northeastern Thai (or Isan, 15 million people), Northern Thai (or Kam Meuang or Lanna, 6 million people), Southern Thai (5 million people), Khorat Thai (400,000 people), and many more variations (http://en.wikipedia.org/wiki/Thai_language). The Thai language is considered a member of the family of Tai languages, the language is used in many parts of the Indochina sub-region including India, southern China, northern Myanmar, Laos, Thai, Cambodia, and North Vietnam.

The Thai script of today has a history going back about 700 years, with gradual changes in the script's shape and writing system evolving over the years. The script was originally derived from the Khmer script in the sixth century. It is generally thought that the Khmer script developed from the Pallava script of India.

Pronunciation of Thai words does not change with their context of use, as each word has a fixed tone. Changing the tone of a syllable (change of the tone mark) may lead to a totally different meaning. Thai verbs do not change their forms as to tense, gender, and singular or plural form. Instead, there are other additional words to help with the meaning for tense, gender, and singular or plural. Basic Thai words are typically monosyllabic. Contemporary Thai makes extensive use of adapted Pali, Sanskrit, English, and Chinese words embedded in day-to-day vocabulary. Some words have been in use long enough that people have forgotten that they originated from other languages.

Thai is written left-to-right, without spaces between words. Each character has only one form, that is, no notion of uppercase and lowercase characters. Some vowels are written before or after the main consonant. Certain vowels, all tone marks, and diacritics are written above or below the main character.

A Thai word is typically formed by the combination of one or more consonants, one vowel (or one composited vowel), none or one tone mark, and optional one or more final consonants to make one syllable. Certain words may be polysyllabic and therefore they may consist of many characters in combination. There are 87 characters used to represent the language as shown in Table 2.



Consonants	44	ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ล ว ศ ษ ส ห ฬ อ ฮ
Vowels	18	ะ ◌ั า ◌ำ ◌ิ ◌ี ◌ึ ◌ื ◌ุ ◌ู เ แ โ ใ ไ ๅ ฤ ฦ
Tone marks	4	◌่ ◌้ ◌๊ ◌๋
Diacritics	5	◌ฺ ◌็ ◌์ ◌ํ ◌๎
Numerals	10	๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙
Other symbols	6	ฯ ฿ ๆ ๏ ๚ ๛
Total	87	

Table 2: The Thai Characters

3.2 Northeastern Thai (Isan)

Northeastern Thai or Isan is a group of Lao dialects spoken in the northern two-thirds of northeastern Thailand, also known as the Isan region. It is spoken by 20 million or so people in Thailand, and 80% of all Lao speakers. The language remains the primary spoken language in households in Isan. The Isan language has unofficial status in Thailand and can be differentiated as a whole from the Lao language of Laos by the increasing use of the Thai grammar, vocabulary and neologisms.

Within Thailand, Isan is considered a regional dialect of Thai, however, outside of Thailand, the language is classified as either its own Lao-Phuthai language due to social and historical reasons or generally as just a distinct subset of the Lao language. Thai, Isan, and Lao are all mutually intelligible to some degree, but Isan is closer to standard Lao than to standard Thai in ordinary speech. Thai, Isan and Lao share most of their basic vocabulary as well as a large corpus of shared Sanskrit, Pali, and Khmer loanwords in academic and high-brow language.

The Isan language was previously written in the ancient Lao alphabet. However, since 1871 (2414 B.E.) The government implemented a policy of using standard Thai in the classroom. Isan today is an unwritten language, but if needed, it is often written in the Thai alphabet, such as in the lyrics of karaoke videos from Isan. The Lao language in Laos continues to be written in Lao alphabet as its official script.

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

Due to the intelligibility between Thai and Isan, and because Isan today is often written in Thai alphabet by transliteration following the Thai spelling rules, therefore the Thai GP agreed that Northeastern Thai or Isan that is written in Thai is naturally covered by the Thai LGR. And if it is written in Lao alphabet, then it should be well covered by the Lao LGR.

3.3 Northern Thai (Lana, or Kam Mueang)

Northern Thai, (Lanna, or Kam Mueang) is the language of the Northern Thai people of Lana, the northern region of Thailand. It is a Tai language closely related to Lao. Northern Thai has approximately six million speakers, most of whom live in Thailand, with a few thousand in northwestern Laos.

Currently, different scripts are used to write Northern Thai. Northern Thai is traditionally written with the Tai Tham script, which in Northern Thai is called Tua Mueang. Tua Mueang is closely related to the old Tai Lue alphabet and the Lao religious alphabets and it is now largely limited to Buddhist temples, where many old sermon manuscripts are still in active use. There is no active production of literature in the traditional alphabet.

Native speakers are presently illiterate in the traditional script therefore, in Thailand, they instead use the Thai script to write the language. Some problems arise when the Thai script is used to write Northern Thai. In particular, the Standard Thai script cannot transcribe all Northern Thai tones. The two falling tones in Northern Thai correspond to a single falling tone in Thai.

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

It is stated in MSR-2 that Lana or Tai Tham are candidate scripts for possible future MSRs (MSR-2 section 3.9 page 11). In the future, Lana could be covered by its own LGR. However, today speakers of these languages are using the Thai Script in transliteration, following the same word formation system. The Thai GP concludes that Northern Thai or Lana that is written in the Thai Script is covered by the Thai LGR.

3.4 Southern Thai (Pak Tai)

Southern Thai, also known as Pak Tai, is a Southwestern Tai language spoken in the fourteen provinces of Southern Thailand as well as by small communities in the northernmost Malaysian states. It is spoken by roughly five million people, and as a second language by the 1.5 million speakers of Kelantan-Pattani Malay and other ethnic groups such as the local Thai Chinese communities, Negritos, and other tribal groups. Most speakers are also fluent in, or understand, the Central Thai dialects.

Southern Thai is mainly a spoken language, although the Thai alphabet is often used in the informal situations when it is written. The words used that are etymologically Thai are often spoken in a reduced and rapid manner, making comprehension by speakers of other varieties difficult. Also, as Southern Thai uses up to seven tones in certain provinces, the tonal distribution is different from other regional varieties of Thai. Additionally, Southern Thai speakers almost always preserve ร as /r/ in contrast to Northern Thai, the Lao-based Isan language, and informal registers of the Standard Thai where it is generally realized as /l/.

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

Similar to Northeastern Thai and Northern Thai, as it is written in transliteration with the Thai Script, the word formation rules are the same and therefore are automatically covered by the Thai LGR.

3.5 Northern Khmer

Northern Khmer, also called Khmer Surin, is the dialect of the Khmer language spoken by approximately 1.4 million Khmer native to the Thai provinces of Surin, Sisaket, Buriram and Roi Et as well as those that have migrated from this region into Cambodia.

Northern Khmer differs from the standard language, based on a dialect of Central Khmer, in the number and variety of vowel phonemes, consonantal distribution, lexicon, grammar, and, most notably, pronunciation of syllable-final /r/, giving Northern Khmer a distinct accent easily recognizable by speakers of other dialects.

Northern Khmer is, for the most part, a spoken language as most speakers are unable to read or write their native tongue (https://en.wikipedia.org/wiki/Northern_Khmer_dialect). In Thailand, Northern Khmer is written in the Thai script.

As many sounds occur in Northern Khmer that would be impossible to write according to the rules of the Thai orthography, a few innovations are necessary such as using ฮ (initial /h/ in Thai) at the end of words to represent syllable-final /h/ and ญ (initial /j/, final /n/ in Thai) to represent Northern Khmer's palatal nasal /ɲ/. Special Thai character diacritics are also sometimes used with the vowels because Northern Khmer has more vowel positions than Thai. These character diacritics are in MSR-2, code points 0E3A PHINTU and 0E4D NIKHAHIT, therefore LGR for Northern Khmer will be included in the Thai LGR.

https://th.wikipedia.org/wiki/ภาษาเขมรถิ่นไทย

3.6 Pattany Malay (Kelantan-Pattany Malay, or Yawi)

Kelantan-Pattani Malay, often referred to in Thailand as Yawi (in Thai) or Jawi (in Patani Malay), and in Kelantan as Bahasa Melayo Kelate (بهاس ملايو كلنتن), is a Malayan language spoken in the Malaysian state of Kelantan and the neighbouring southernmost provinces of Thailand. It is the primary spoken language of Thai Malays, but is also used as a lingua franca by ethnic Southern Thais in rural areas, Muslim and non-Muslim, and the samsam, a mostly Thai-speaking population of mixed Malay and Thai ancestry.

Pattany Malay is originally written in the Jawi alphabet, based on the Arabic script, which is where the name "Yawi/Jawi" for the language comes from. Today, Pattani Malay itself is generally not a written language, though it is sometimes written in informal settings. The general population of Malay speakers in both Malaysia and Indonesia now use the Latin script, known in Malay as Rumi (رومي), for daily communication.

A phonetic rendering of Pattani Malay in the Thai alphabet has been introduced, but it has not met with much success, due to the socio-religious significance of Jawi to Muslim Malays, as well as because of numerous inconsistencies and inaccuracies.

As Pattany Malay is mostly written in Arabic or Latin Script the Thai GP considered that including Pattany Malay code points in the Thai LGR might be premature and therefore excluded it from consideration.

3.7 Summary of Languages covered by the Thai LGR

The Thai GP set the principle for language selection to any language that is written or may be written in the Thai Script which is still actively used and has been reported as having high-usage population (not less than 1 million). There are 6 languages eligible by this criterion: Thai, Northeastern Thai, Northern Thai, Southern Thai, North Khmer, and Pattani Malay.

Thai is the de fact to national language with a fully developed writing system and it will be the main script to consider in Thai LGR, which will also naturally cover Northeastern Thai, Northern Thai, Southern Thai and Northern Khmer (in Thai Script).

However, the Thai GP will exclude Pattany Malay from the Thai LGR because the dominant writing system for it is either Arabic-based (Yawi) or Latin-based (Rumi).

4 Overall Development Process and Methodology

The generation panel started the work from October 2015 and had been discussing to develop the Thai Label Generation Rules proposal through various types of meetings such as regular formal meetings at the early stage of the proposal development. In addition, the panel has been working via mailing list as a normal basis, and it holds teleconference after reaching each milestone. As the majority of the panel members are in Bangkok, any face-to-face meetings were organized by ETDA.

Furthermore, the panel held an open public consultation before submitting the proposal to ICANN. The purpose of public consultation aimed for gathering feedback on the work from the larger community and experts. Experts were invited from Thai Internet Governance Forum such as, but not limited to, internet users, IT companies, internet service providers, universities and media. The feedback from this consultation was used in finalizing of the proposal for submission.

5 Code Point Repertoire

The Thai GP takes code points shortlisted in MSR-2 as a starting point for Thai Script analysis for Root Zone Label Generation Rules. The Thai GP makes reference to the Thai Script writing system from the Royal Institute of Thailand and also refers to various standards for using Thai in computers. These standards from the Thai Industrial Standard Institute (TISI) include, TIS 620 series – Standard for Thai Character codes for Computers, TIS 820 series – Layout of Thai Character Keys on Computer Keyboard, and TIS 1566 – Thai Input/ Output Methods for Computers.

The Thai Script is an abugida script, in which consonant–vowel sequences are written as a unit: each unit is based on a consonant letter; vowels, tone marks, or diacritic notation are secondary. It is written with the combining marks stacked above or below the base consonant, like diacritics in European languages. However, although the concepts are quite similar, the implementations are significantly different.

First, there are too many possible combinations of base consonants and combining marks in Thai to be enumerated like Latin accents in the ISO/IEC 8859 series standard for 8-bit character encoding. Therefore, base characters and combining characters are encoded separately, rather than pre-combined.

Second, Thai combining marks are classified into upper or lower vowels, tone marks, and other diacritics. The base consonant can be combined with up to two combining marks, that is, zero or one upper or lower vowel and zero or one tone mark or diacritic. The upper/lower vowel, if present, is always attached to the consonant before the tone/diacritic.

A unique characteristic of a script that allows both upper and lower position is that the different input order of upper and lower marks can produce a visually identical word, while the stored codes are different. This can lead to problematic issues, such as, failures in string matching or confusing the output method.

To solve this problem, in 1990, the Thai API Consortium, a group of Thailand software developers led by Thaweesak Koanantakool, drafted a common specification for computer handling of the Thai I/O method. The work was funded by NECTEC and was published in 1998 as the WTT 2.0 specification. The WTT 2.0 specification defines the canonical order of Thai character strings. WTT 2.0 compliance requires that certain input-sequence rules must be met, and most (if not all) input syntactic errors are eliminated at the time of data entry. The WTT2.0 later become the national standard “TIS 1566 – Thai Input/ Output Methods for Computers”.

A known implementation of the WTT-based Thai input method covers Microsoft DOS and Windows (all versions), the Thai language Environment for Solaris and a few other platforms. Even though there is a small chance that any WTT2.0 non-compliant string will reach the point of being considered for a label in root zone, it is still worth stating the sequence rule explicitly in a later section of this document.

This section includes a summary of analysis of the code points, based on which the repertoire has been selected.

According to the National Standard TIS 1566, excluding the numbers and control characters, the Thai Script characters are classified as shown in Table 3.

Type	Numbers	Subtyp	Characters
Consonants	44	CONS: consonant	ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ล ว ศ ษ ส ห ฬ อ ฮ
Vowels	5	LV: leading vowel	เ แ โ ใ ไ

Type	Numbers	Subtyp	Characters
	6	FV1: ordinary forwarding vowel	ะ า ◌ำ
		FV2: dependent forwarding vowel	ๅ
		FV3: special forwarding vowel	ฤ ฦ
	2	BV: below vowel	◌ุ ◌ู
	5	AV: above vowel	◌ั ◌ิ ◌ี ◌ึ ◌ื
Tone Marks	4	TONE: tone mark	◌่ ◌้ ◌๊ ◌๋
Diacritics	5	AD: above diacritic	◌็ ◌์ ◌ํ ◌๎
		BD: below diacritic	◌ฺ
Other Symbols and Marks	6	NON: non-decomposable	ฯ ฿ ๏ ๆ ๚ ๛

Table 3: Classification of Thai characters based on TIS-1566

5.1 Consonants

Thai consonants can be classified as plosives (stops), non-plosives, sibilants, and voiced ''h.'' Table 3 shows the subtypes associated with each character. Table 4 classifies the consonants, showing the associated glyphs in the International Phonetic Alphabet (created by the International Phonetic Association, IPA)

In each cell below, the first line indicates International Phonetic Alphabet (IPA), the second indicates the Thai characters in initial position (several letters appearing in the same box have identical or nearly identical pronunciation).

		Labial	Alveolar	Palatal	Velar	Glottal
Nasal		[m] ม	[n] ณ, น		[ŋ] ง	
Plosive	voiced	[b] บ	[d] ฎ, ด			
	tenuis	[p] ป	[t] ฏ, ต	[tɕ] จ	[k] ก	[ʔ] อ
	aspirated	[pʰ] ผ, พ, ภ	[tʰ] ฐ, ฑ, ฒ, ถ, ท, ธ	[tɕʰ] ฉ, ช, ฌ	[kʰ] ข, ฃ*, ค, ฅ*, ฆ	
Fricative		[f] ฝ, ฟ	[s] ซ, ศ, ษ, ส			[h] ห, ฮ
Approximant			[l] ล, ฬ	[j] ญ, ย	[w] ว	

Trill		[r] ร			

Table 4: Classification of Thai consonants. Each cell consists of the character's IPA symbol in square brackets and the Thai letters that have the given pronunciation.

Note:

- U+0E2D (O ANG, อ), a plosive glottal, is considered a zero-consonant and can be used to write stand-alone vowels.
- U+0E03 (KHO KHUAT, ฃ) is now rarely used. There are currently no words using KHO KHUAT in the Thai language according to the Royal Institute Dictionary of 1999, the official standard current dictionary of the Thai language. However, KHO KHUAT still has an entry in most dictionaries stating that it is obsolete, and is included on alphabet charts to preserve the traditional count of 44 Thai consonants.
- U+0E05 (KHO KHON, ฅ) is also no longer exists in Modern Thai. Similarly, it remains in dictionaries, preserving the traditional count of 44 letters in the Thai alphabet.
- The main reason for declining usage was that when the first Thai typewriter was developed by Edwin Hunter McFarland in 1892, there was simply no space for all characters, thus KHO KHUAT and KHO KHON were just left out as they can be replaced phonetically by U+0E02 (KHO KHAI) and U+0E04 (KHO KHWAI) respectively.
- Even though they are no longer exist in Modern Thai, since 2005, there has been a small movement to bring these two characters back in use. One such publisher is Butterfly Book House (สำนักพิมพ์ผีเสื้อ), which publishes children's literature by Thai authors and Thai translations of foreign authors, such as Roald Dahl. In its books, words like ขวด and คน are spelled ฃวด and ฅน, despite the fact this appears "conservative". As another example, it was used in the movie title "ฅนไฟบิน" (in English: Dynamite Warrior, 2006)

Considering these points above, even though the two code points are no longer used in the dictionary words, it is still possible to use them in names or brands. Moreover, the Standard Thai Keyboard Layout includes these two characters. Therefore, Thai GP decided to include all 44 code points for consonants.

5.2 Vowels

The 18 vowel symbols pronounced after a consonant are non-sequential in writing: they can be located before, after, above or below the consonant, or in a combination of these positions. The symbols are listed in Table 5, with the associated position type and Unicode values.

#	Vowel	Unicode Value	Name	Type	Position relative to the main consonant
1	ะ	0E30	SARA A	FV1	Follow
2	◌ั	0E31	MAI HAN-AKAT	AV	Above

#	Vowel	Unicode Value	Name	Type	Position relative to the main consonant
3	า	0E32	SARA AA	FV1	Follow
4	◌ำ	0E33	SARA AM	FV1	Follow
5	◌ิ	0E34	SARA I	AV	Above
6	◌ี	0E35	SARA II	AV	Above
7	◌ึ	0E36	SARA UE	AV	Above
8	◌ื	0E37	SARA UEE	AV	Above
9	◌ุ	0E38	SARA U	BV	Below
10	◌ู	0E39	SARA UU	BV	Below
11	เ	0E40	SARA E	LV	Leading
12	แ	0E41	SARA AE	LV	Leading
13	โ	0E42	SARA O	LV	Leading
14	ใ	0E43	SARA AI MAIMUAN	LV	Leading
15	ไ	0E44	SARA AI MAIMALAI	LV	Leading
16	ๅ	0E45	LAKKHANGYAO	FV2	Follow
17	ฤ	0E24	RU	FV3	Follow
18	ฦ	0E26	LU	FV3	Follow

Table 5: Eighteen vowel symbols in Thai

Note:

- U+0E33 (SARA AM) is not allowed in IDNA2008 and will be decomposed into U+0E4D (NIKHAHIT) and U+0E32 (SARA AA).
- U+0E45 (LAKKHANGYAO) has the same phonetic function as SARA AA, but always follows the special vowel 0E24 (RU) or 0E26 (LU), with no exceptions. Following the [Guideline] this type of requirement can be satisfied by removing the code point U+0E45 (THAI CHARACTER LAKKHANGYAO, ๅ) from the code point repertoire, and, instead, adding two code point sequences, <U+0E24 U+0E45>, ฤๅ and <U+0E24 U+0E45> ฦๅ.
- The 18 vowels symbols, together with three consonants—ย, ว, and อ—are used in combination to create 32 vowels for Thai, as shown in Table 6.

As discussed above, The Thai GP decided to exclude U+0E45 (LAKKHANGYAO) from the code point repertoire, and, instead add two code point sequences: as U+0E24 U+0E45>, ฤๅ and <U+0E24 U+0E45> ฦๅ.

Important note:

This code point analysis is intended for the purpose of LGR only. Excluding some code points is not intended, by any means, to undermine the recognition of U+0E45 (LAKKHANGYAO). It should be strictly recognized as an active Thai vowel, but outside the context of LGR.

	Front		Back			
	Unrounded		Unrounded		Rounded	
	Short	Long	Short	Long	Short	Long
Close	-ิ /i/	-ี /iː/	-ึ /ɯ/	-ื /ɯː/	-ุ /u/	-ู /uː/
Close-mid	เ-ะ /e/	เ- /eː/	เ-อะ /ɤ/	เ-อ /ɤː/	โ-ะ /o/	โ- /oː/
Open-mid	แ-ะ /ɛ/	แ- /ɛː/	-	-	เ-าะ /ɔ/	-อ /ɔː/
Open	-	-	-ะ, -ั /a/	-า /aː/		

(a) 18 Monophthongs

Thai	-ัวะ	-ัว	เ-ียะ	เ-ีย	เ-ือะ	เ-ือ	ใ-	ไ-	เ-า
IPA	/ua/	/uːa/	/ia/	/iːa/	/ɯa/	/ɯːa/	/aj/	/aj/	/aw/

(b) 9 Diphthongs

Thai	-ำ	ฤ	ฤๅ	ฦ	ฦๅ
IPA	/am/	/rɯ/, /ri/, /rɤː/	/rɯː/	/lɯ/	/lɯː/

(c) 5 Semi-vowels

Table 6: Thirty-two vowels in Thai: (a) 18 monophthongs, (b) 9 diphthongs, and (c) 5 semi-vowels

5.3 Tone marks

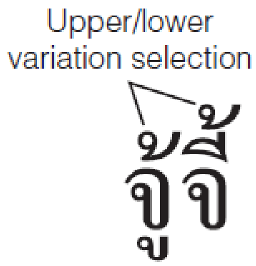
There are five phonemic tones: mid, low, falling, high, and rising. These 5 tones are represented by 4 tone marks and no mark. It is important to pronounce each syllable in the proper lexical tone; different tones create entirely different meanings. The tone rules for pronunciation consider the specific consonants, vowels, and tone marks in a syllable to determine the tone with which the syllable must be pronounced. These rules are not relevant to IDNs are thus not included in this document.

For the writing system, tone marks, if any, are always placed above the final onset consonant of the syllable with these display rules

- Tone marks are placed above the final onset consonant of the syllable (Table 7)
- Select the lower variation for top most mark in the absence of an upper vowel (Picture 1)

#	Tone Mark	Unicode Value	Name	Example	Gloss
1	◌่	0E48	MAI EK	เก่า	Old
2	◌้	0E49	MAI THO	เก้า	Nine
3	◌๊	0E4A	MAI TRI	เกี๊ยว	Dumpling
4	◌๋	0E4B	MAI CHATTAWA	เดี๋ยว	Just now

Table 7: Example of tone mark position above the final onset consonant of the syllable



Picture 1: In the absence of an upper vowel, the glyph for the lower variation of the tone mark is selected

5.4 Diacritics

There are 5 diacritic symbols with differences in their frequency and purpose of usage.

- U+0E47 (MAITAIKHU) and U+0E4C (THANTHAKHAT) are commonly used in everyday communication.

U+0E3A (PHINTHU) is normally limited to use in Buddhist temples, where many old sermon Pali manuscripts are still in active use. However, to support other languages in Thailand, those that use the Thai Script for transliteration, U+0E3A (PHINTHU) is occasionally used, for example, the word “Allah”, the Arabic word referring to God which also appears in Pattany Malay. That word can be transliterated as ”อัลลอฮ์”, “อัลลอฮฺ”, “อัลลอหฺ” or “อัลเลาะห์”.

- Therefore, the Thai GP decided to include U+0E3A (PHINTHU)
- U+0E4D (NIKHAHIT) is also limited to Buddhist temples, where many old sermon Pali manuscripts are still in active use. It is used in decomposing U+0E33 (SARA AM, ◌ำ) which is commonly used. It is necessary to be included
- U+0E4E (YAMAKKAN) is rarely used in the modern Thai or even the old sermon Pali manuscript as more common to replace with U+0E3A (PHINTHU)

The position of diacritics is shown in Table 7. The Thai GP decided to exclude U+0E4E (YAMAKKAN) from the Thai LGR repertoire as it is rarely used. Moreover, excluding U+0E4E (YAMAKKAN) will also reduce the chance of confusion between U+0E4E (YAMAKKAN) and U+0E4C (THANTHAKHAT). Both look similar, and both are always placed at the same position in the word cell, and are normally displayed in a small size.

#	Diacritics	Unicode Value	Name	Type	Position relative to the main consonant
1	◌ฺ	0E3A	PHINTHU	BD	Below
2	◌็	0E47	MAITAIKHU	AD2	Above
3	◌์	0E4C	THANTHAKHAT	AD1	Above
4	◌ํ	0E4D	NIKHAHIT	AD1	Above
5	◌๎	0E4E	YAMAKKAN	AD3	Above

Table 7: The position of diacritics in the Thai Script

5.5 Numerals

The Thai digits from U+0E50 (Zero, ๐) to U+0E59 (Nine, ๙) are treated in the same fashion as digits in other scripts in that they are not allowed in the root zone. They are out of scope for this proposal.

5.6 Other Symbols and Marks

There are six other Thai symbols and marks that are used as function symbols. IDNA2008 already disallowed four of them and MSR-2 also excluded one. See Table 8.

U+0E46 (MAIYAMOK) is a repetition mark, for example “บ้านบ้าน” can be written as “บ้าน ๆ” which has the same meaning (home-home).

The discussion on including the repetition mark with the Thai LGR is not yet conclusive within the Thai GP. Some members would like to include the mark because it is commonly used and friendly for marketing as it shortens a string, and currently it is allowed in the second level IDN tables, while other members are considering not including it because it could cause confusion for consumers.

This discussion may take some time to conclude, due to the conservatism principle. Thai GP recommends not including it in the repertoire at this time.

However, the discussion of this issue will be continued within the Thai GP and among the public multi-stakeholders in Thailand. If the resolution is to include U+0E46 (MAIYAMOK) into the repertoire, there will be a detailed amendment to this proposal.

#	Vowel	Unicode Value	Name	Function	IDNA 2008	MSR-2	Thai LGR
1	ฯ	0E2F	PAIYANNOI	ellipsis, abbreviation	PVALID	excluded	■

#	Vowel	Unicode Value	Name	Function	IDNA 2008	MSR-2	Thai LGR
2	฿	0E3F	SYMBOL BAHT	currency	DISALLOWED		
3	ๆ	0E46	MAIYAMOK	repetition	PVALID		excluded
4	๏	0E4F	FONGMAN	used as a bullet	DISALLOWED		
5	๚	0E5A	ANGKHANKHU	used to mark end of long sections	DISALLOWED		
6	๛	0E5B	KHOMUT	used to mark end of chapter or document	DISALLOWED		

Table 8: Analysis of other symbols and Marks in the Thai Script

5.7 Summary of code point repertoire included and excluded

Starting from the 71 code points in MSR-2 for the Thai Script, the Thai GP considered 3 code points to be excluded (Table 10). One code point (U+0E45 LAKKHANGYAO) is only supported when it occurs in one of two additional code point sequences defined in this LGR – it can be part of a label but only if used in the context of one of these sequences. The repertoire includes 68 code points and three code point sequences (Table 9).

All code points are referenced in Thai Industrial Standard (TIS) 1566 – Thai Input/ Output Methods for Computers, which contains the Thai Script writing system from Royal Institute of Thailand. This standard in turn cites various standards for using Thai in computers from the Thai Industrial Standard Institute, such as TIS 620 series – Standard for Thai Character codes for Computers and TIS 820 series – Layout of Thai Character Keys on Computer Keyboard.

See chart for code points in Appendix 1.

5.7.1 Code point repertoire included

All code points in the repertoire are from the Thai language, which has an EGIDS score of 1, indicating that the language is used in education, work, mass media, and government at the national level.

#	Unicode Code Point	Glyph	Unicode Code Point Name	Unicode General Category[1]	Category/Tag[2]
1	0E01	ก	THAI CHARACTER KO KAI	Lo	cons
2	0E02	ข	THAI CHARACTER KHO KHAI	Lo	cons
3	0E03	ฃ	THAI CHARACTER KHO KHUAT	Lo	cons

[1] http://unicode.org/cldr/utility/character.jsp

[2] Category by TIS-1566 Thai Script Characters are classification, and additional tag for WLE rule purpose

#	Unicode Code Point	Glyph	Unicode Code Point Name	Unicode General Category[1]	Category/Tag[2]
4	0E04	ค	THAI CHARACTER KHO KHWAI	Lo	cons
5	0E05	ฅ	THAI CHARACTER KHO KHON	Lo	cons
6	0E06	ฆ	THAI CHARACTER KHO RAKHANG	Lo	cons
7	0E07	ง	THAI CHARACTER NGO NGU	Lo	cons
8	0E08	จ	THAI CHARACTER CHO CHAN	Lo	cons
9	0E09	ฉ	THAI CHARACTER CHO CHING	Lo	cons
10	0E0A	ช	THAI CHARACTER CHO CHANG	Lo	cons
11	0E0B	ซ	THAI CHARACTER SO SO	Lo	cons
12	0E0C	ฌ	THAI CHARACTER CHO CHOE	Lo	cons
13	0E0D	ญ	THAI CHARACTER YO YING	Lo	cons
14	0E0E	ฎ	THAI CHARACTER DO CHADA	Lo	cons
15	0E0F	ฏ	THAI CHARACTER TO PATAK	Lo	cons
16	0E10	ฐ	THAI CHARACTER THO THAN	Lo	cons
17	0E11	ฑ	THAI CHARACTER THO NANGMONTHO	Lo	cons
18	0E12	ฒ	THAI CHARACTER THO PHUTHAO	Lo	cons
19	0E13	ณ	THAI CHARACTER NO NEN	Lo	cons
20	0E14	ด	THAI CHARACTER DO DEK	Lo	cons
21	0E15	ต	THAI CHARACTER TO TAO	Lo	cons
22	0E16	ถ	THAI CHARACTER THO THUNG	Lo	cons
23	0E17	ท	THAI CHARACTER THO THAHAN	Lo	cons
24	0E18	ธ	THAI CHARACTER THO THONG	Lo	cons
25	0E19	น	THAI CHARACTER NO NU	Lo	cons
26	0E1A	บ	THAI CHARACTER BO BAIMAI	Lo	cons
27	0E1B	ป	THAI CHARACTER PO PLA	Lo	cons

#	Unicode Code Point	Glyph	Unicode Code Point Name	Unicode General Category[1]	Category/Tag[2]
28	0E1C	ผ	THAI CHARACTER PHO PHUNG	Lo	cons
29	0E1D	ฝ	THAI CHARACTER FO FA	Lo	cons
30	0E1E	พ	THAI CHARACTER PHO PHAN	Lo	cons
31	0E1F	ฟ	THAI CHARACTER FO FAN	Lo	cons
32	0E20	ภ	THAI CHARACTER PHO SAMPHAO	Lo	cons
33	0E21	ม	THAI CHARACTER MO MA	Lo	cons
34	0E22	ย	THAI CHARACTER YO YAK	Lo	cons
35	0E23	ร	THAI CHARACTER RO RUA	Lo	cons
36	0E24	ฤ	THAI CHARACTER RU	Lo	fv3
37	0E25	ล	THAI CHARACTER LO LING	Lo	cons
38	0E26	ฦ	THAI CHARACTER LU	Lo	fv3
39	0E27	ว	THAI CHARACTER WO WAEN	Lo	cons
40	0E28	ศ	THAI CHARACTER SO SALA	Lo	cons
41	0E29	ษ	THAI CHARACTER SO RUSI	Lo	cons
42	0E2A	ส	THAI CHARACTER SO SUA	Lo	cons
43	0E2B	ห	THAI CHARACTER HO HIP	Lo	cons
44	0E2C	ฬ	THAI CHARACTER LO CHULA	Lo	cons
45	0E2D	อ	THAI CHARACTER O ANG	Lo	cons
46	0E2E	ฮ	THAI CHARACTER HO NOKHUK	Lo	cons
47	0E30	ะ	THAI CHARACTER SARA A	Lo	fv1
48	0E31	◌ั	THAI CHARACTER MAI HAN-AKAT	Mn	av
49	0E32	า	THAI CHARACTER SARA AA	Lo	fv1, sara-aa
50	0E34	◌ิ	THAI CHARACTER SARA I	Mn	av
51	0E35	◌ี	THAI CHARACTER SARA II	Mn	av

#	Unicode Code Point	Glyph	Unicode Code Point Name	Unicode General Category[1]	Category/Tag[2]
52	0E36	◌ึ	THAI CHARACTER SARA UE	Mn	av
53	0E37	◌ื	THAI CHARACTER SARA UEE	Mn	av
54	0E38	◌ุ	THAI CHARACTER SARA U	Mn	bv
55	0E39	◌ู	THAI CHARACTER SARA UU	Mn	bv
56	0E3A	◌ฺ	THAI CHARACTER PHINTHU	Mn	bd
57	0E40	เ	THAI CHARACTER SARA E	Lo	lv
58	0E41	แ	THAI CHARACTER SARA AE	Lo	lv
59	0E42	โ	THAI CHARACTER SARA O	Lo	lv
60	0E43	ใ	THAI CHARACTER SARA AI MAIMUAN	Lo	lv
61	0E44	ไ	THAI CHARACTER SARA AI MAIMALAI	Lo	lv
62	0E47	◌็	THAI CHARACTER MAITAIKHU	Mn	ad, maitaikhu
63	0E48	◌่	THAI CHARACTER MAI EK	Mn	tone
64	0E49	◌้	THAI CHARACTER MAI THO	Mn	tone
65	0E4A	◌๊	THAI CHARACTER MAI TRI	Mn	tone
66	0E4B	◌๋	THAI CHARACTER MAI CHATTAWA	Mn	tone
67	0E4C	◌์	THAI CHARACTER THANTHAKHAT	Mn	ad, thanthakhat
68	0E4D	◌ํ	THAI CHARACTER NIKHAHIT	Mn	ad, nikhahit
69	0E24 + 0E45	ฤๅ	THAI CHARACTER RU + THAI CHARACTER LAKKHANGYAO		fv2
70	0E26 + 0E45	ฦๅ	THAI CHARACTER LU + THAI CHARACTER LAKKHANGYAO		fv2
71	0E4D + 0E32	◌ํา	THAI CHARACTER NIKHAHIT + THAI CHARACTER SARA AA		fv1

Table 9: Code point repertoire included

5.7.2 Code point repertoire excluded

#	Unicode Code Point	Glyph	Unicode Code Point Name	Unicode General Category	Category
1	0E45	ๅ	THAI CHARACTER LAKKHANGYAO	Lo	fv2
2	0E46	ๆ	THAI CHARACTER MAIYAMOK	Lm	non
3	0E4E	◌๎	THAI CHARACTER YAMAKKAN	Mn	ad

Table 10: Code points excluded from repertoire

6 Variants

6.1 Script-Internal homoglyphs for Thai

The Thai GP considers two code points as variants if they are visually the same or very similar to each other. Based on this, there are three possible variant cases to be discussed.

- **U+0E40 (THAI CHARACTER SARA E, เ) and U+0E41 (THAI CHARACTER SARA AE, แ)**

  U+0E41 is a digraph of <U+0E40, U+0E40>, that is, a sequence of two instance of THAI CHARACTER SARA E is visually the same as THAI CHARACTER SARA AE. So the two sequences <U+0E41> and <U+0E40, U+0E40> are variants.

  Both of them are leading vowels, which will appear before the onset consonant. The Thai writing system does not allow two consecutive leading vowels in a word and this rule will be proposed in WLE. Therefore, the Thai GP has discussed and decided that there is no separate need to block these variants from each other.

- **U+0E32 (THAI CHARACTER SARA AA, า) and U+0E45 (THAI CHARACTER LAKKHANGYAO, ๅ)**

  These are visually similar, the difference between them being that the U+0E45 has a longer downward stem. However, the usage is very different and discrete from each other. U+0E32 (THAI CHARACTER SARA AA, า) is a vowel that is eligible to follow after any of the 44 consonants, while U+0E45 (THAI CHARACTER LAKKHANGYAO, ๅ) is eligible to follow only two special vowels, the U+0E24 (THAI CHARACTER RU, ฤ) to become ฤๅ, and the U+0E26 (THAI CHARACTER LU, ฦ) to become ฦๅ.

  This requirement is already covered by removing U+0E45 (THAI CHARACTER LAKKHANGYAO, ๅ) from the code points, and adding two extra code point sequences, <U+0E24 U+0E45>, ฤๅ and <U+0E24 U+0E45> ฦๅ.

- **U+0E33 (THAI CHARACTER SARA AM, ◌ำ) and U+0E4D (THAI CHARACTER NIKHAHIT, ◌ํ) + U+0E32 (THAI CHARACTER SARA AA, า)**

  U+0E33 and <U+0E4D, U+0E32> are variants. However, U+0E33 (THAI CHARACTER SARA AM, ◌ำ) is already excluded from IDNA2008, therefore it is out of scope of this proposal.

In conclusion, no blocking script-internal variants are proposed for the Thai Script.

6.2 Cross Script homoglyphs

The Thai GP has analyzed Lao, Khmer and Myanmar scripts, which are historically related scripts and for which some cross-script homoglyphs might exist. Some consonants and vowels are similar especially between the Thai Script and the Lao Script, but they are not homoglyphs to the degree that would warrant considering them as cross-script variants.

The Thai GP also had a chance to discuss this issue with the Khmer GP and the Lao GP who both agree that there are no cross-script variants. The samples of similar scripts are listed in Appendix B.

In conclusion, the Thai GP has analyzed Thai consonants and vowels compared to Khmer, Lao and Myanmar consonants and vowels. Some of consonants and vowels are very similar as discussed above. However, as they look different especially when in combined form in a label, there are no cross-script variants proposed.

7 Whole Label Evaluation Rules (WLE)

Thai is a complex script in which a sequence of code points creates a character cluster in a cell, and only a subset of all possible code point sequences would ever be expected to occur. The WLE rules in this LGR are used to limit the context in which certain code points or marks may appear, so that they fall in the range expected (and supported) by typical rendering engines, but they are not intended to enforce 'spelling-rules'.

Using simple generalized WLE Rules will also allow the other language users to be able to input a string in their language using the Thai Script without any spelling rules limitation, while still maintaining the consistent behaviour of rendering engines.

All the default rules in MSR-2 also apply to the Thai Script, such as a label cannot start with a combining mark (General Category = Mark, Non-spacing (Mn)). In addition to the default rules, there are some restrictions on label-level to construct the Thai Script label described below.

7.1 No leading combining mark

A label cannot start with a combining mark. This applies to those code points with the General Category of Mn and Mc in the repertoire table (see also Section 7.7).

7.2 Every leading vowel must precede a consonant

Every leading-vowel is a dependent vowel that cannot stand alone and needs at least one following consonant to form a label. It also cannot be followed by another vowel, a diacritic, or a tone-mark.

This rule will also solve the variant issue that U+0E41 is a digraph of <U+0E40, U+0E40>, mentioned in section 6.1, as two consecutive leading-vowels are not eligible.

7.3 Code Points that must follow a consonant

There are subsets of vowels and diacritics that cannot stand alone and need to follow a consonant. They are:

- above-vowel,
- below-vowel,
- below-diacritic
- the code point U+0E47 (THAI CHARACTER MAITAIKHU, ◌็)

This rule will prevent the case of too many code points at the same position (above or below) in a cell, which can cause an unexpected rendering. And these rules will also eliminate the possibility of double vowels which create non-readable label because of missing onset consonant.

7.4 Context of MAI HAN-AKAT

The code point U+0E31 (THAI CHARACTER MAI HAN-AKAT, ◌ั) is a vowel that always occurs between a consonant and either a tone or a consonant.

7.5 Context of SARA-A

The code point U+0E30 (THAI CHARACTER SARA A, ะ) is a vowel that can follow a consonant or a tone or the vowel code point U+0E32 (THAI CHARACTER SARA AA, า).

Normally, a vowel cannot follow another vowel except in two cases:

- U+0E45 (THAI CHARACTER LAKKHANGYAO, ๅ) can follow only two special vowels: U+0E24 (THAI CHARACTER RU, ฤ) and U+0E26 (THAI CHARACTER LU, ฦ)
- U+0E30 (THAI CHARACTER SARA A, ะ) can follow U+0E32 (THAI CHARACTER SARA AA, า)

The former case is handled by adding the two possible code point sequences into the repertoire. However, the latter case is impossible to handle by enumerating code point sequences as there are too many possible combinations. U+0E30 (THAI CHARACTER SARA A, ะ) is an active vowel that can follow any of the 44 consonants, four tones, and the vowel U+0E32 (THAI CHARACTER SARA AA, า) and all are common and actively used in the Thai Script. Therefore, the latter case is handled by this WLE rule.

7.6 Context of SARA-AA

The code point U+0E32 (THAI CHARACTER SARA AA, า) is a vowel that can follow a consonant or a tone.

7.7 Context of tone mark

Every tone-mark always stays at the topmost position of a cluster. It could be above a consonant or above an above-vowel. However, it cannot follow another tone-mark or an above diacritic (MAITAIKHU, THANTHAKHAT) as they could collide at the top most position and can cause a non-predictable rendered label.

The code point U+0E4D (NIKHAHIT, ◌ํ) is an above-diacritic that can follow a tone mark. This is normally wrong for the writing system, but it is allowed in this LGR because of the code point U+0E33 (THAI CHARACTER SARA AM, ◌ำ). The latter code point is excluded from IDNA2008 and will instead be decomposed into 2 vowels, U+0E4D (THAI CHARACTER NIKHAHIT, ◌ํ) and U+0E32 (THAI CHARACTER SARA AA, า). These two code points have been added as a sequence to the repertoire so that, a tone mark can be followed by a NIKHAHIT as part of SARA AM.

A tone-mark is normally used as the closure of a cluster; therefore, it cannot follow a leading-vowel nor below diacritic which will create a non-readable label because of the missing onset consonant.

As tone-marks are non-spacing marks, please note that the rule that a tone-mark cannot be at the label’s starting position is already covered by the default WLE rules.

7.8 Context of above-diacritic

In a cluster, there is only one above-diacritic allowed. Therefore, an above-diacritic must not follow another above-diacritic; otherwise they will collide at the top most position and can cause a non-predictable rendering of the label.

Also, as the diacritics are non-spacing marks, the rule that a diacritic cannot be at the label’s starting position is already covered by the default WLE rules.

8 Contributors

List of advisory committees.

Name	Designation	Organization
Dr. Thaweesak Koanantakool	President of the National Science and Technology Development Agency (NSTDA)	National Science and Technology Development Agency (NSTDA)
Mrs. Surangkana Wayuparb	Executive Director, CEO	Electronic Transactions Development Agency (Public Organization)
Dr. Virach Sornlertlamvanich	Lecturer	Sirindhorn International Institute of Technology

List of panel members.

Name	Designation	Organization	Relevant experience
Mr. Wanawit Ahkuputra (Chair)	Deputy Executive Director of ETDA, Vice-Chair of ICANN Government Advisory Committee	Electronic Transactions Development Agency (Public Organization)	Develop on ICT policy and software standard
Ms. Pitinan Kooarmornpatana (Secretary)	Director of office of Information Technology Infrastructure	Electronic Transactions Development Agency (Public Organization)	DNS/IDNS/UNICODE expert
Ms. Ubolthip Sethakaset	Specialist	Electronic Transactions Development Agency (Public Organization)	Expert in the Thai-script writing system for the Thai language
Mr. Panus Na Nakorn	Specialist	Electronic Transactions Development Agency (Public Organization)	Policy expert
Mr. Jitti Kunphruk	Senior Innovation Analyst	Electronic Transactions Development Agency (Public Organization)	Software development domain and related to the Thai language.
Mr. Pakpoom Tripatana	Director	Thai Name Server Co., Ltd.	ccTLD/IDN ccTLD registry
Mrs. Pensri Arunwatanamongkol	Director	Dot Arai Co., Ltd.	ICANN accredited registrar
Mr. Arthit Suriyawongkul	Director	Foundation for Internet and Civic Culture	Representative of community on internet governance
Dr. Yunyong Teng-amnuay	Board of Director	Thai Network Information Center Foundation	DNS/IDNS/ccTLD expert
Mr. Ome Sivadith	National Technology Officer	Microsoft (Thailand) Limited	Policy expert

Name	Designation	Organization	Relevant experience
Mr. Worapon Pitayaphongpat	Senior Software Engineer Manager	Microsoft (Thailand) Limited	Software development domain and related to the Thai language in Windows operating system
Ms. Monthika Boriboon	Senior Researcher	National Electronics and Computer Technology Center	Thai Linguistics
Dr. Thepchai Supnithi	Head of Language and Semantic Technology Laboratory	National Electronics and Computer Technology Center	Computational Linguistics
Mr. Wichai Termwuttipreecha	Founding member of Mozilla Thai localization team	Mozilla Thailand Community (Mhafai.com)	Computational Linguistics
Mr. Theppitak Karoonboonyanan	Senior researcher		Thai Linguistics

ICANN Staff

- Dr. Sarmad Hussain

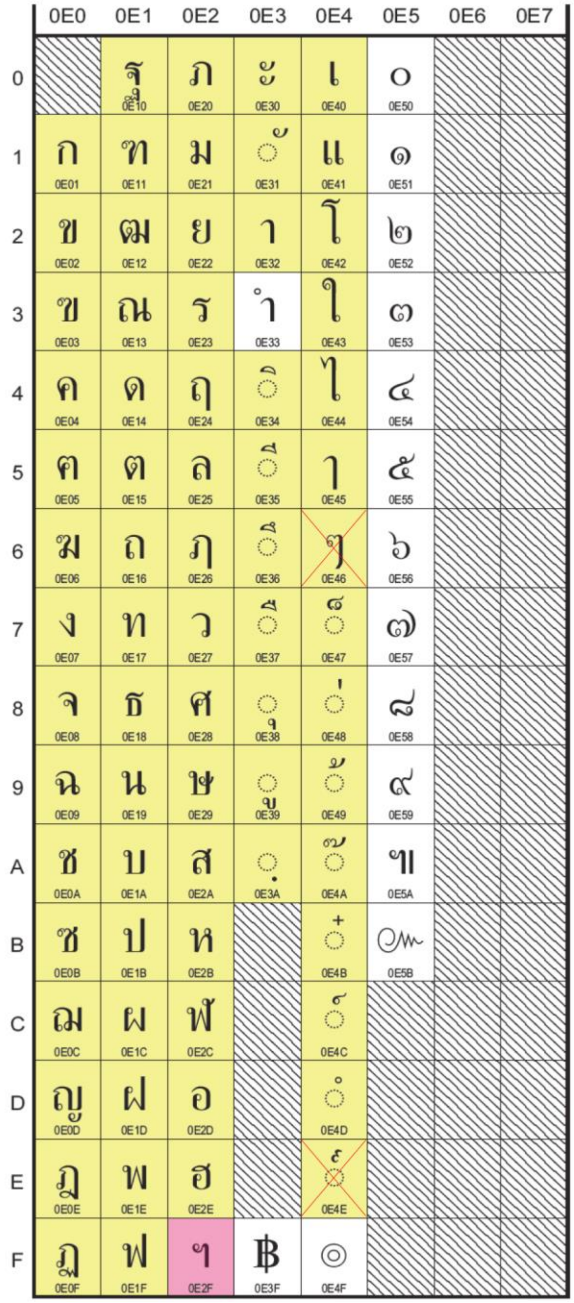
9 References

- Guidelines for Designing Script-Specific Label Generation Rules (LGR) for the Root Zone https//:www.icann.org/news/announcement320150427en
- RFC5892, The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)
- https//:tools.ietf.org/html/rfc5892
- RFC5894,Background, Explanation, and Rationale
- https://tools.ietf.org/html/rfc5894
- RFC5895,Mapping Characters
https//:tools.ietf.org/html/rfc5895
- Representing Label Generation Rulesets Using XML (RFC 7940)
https://tools.ietf.org/html/rfc7940
- Maximal Starting Repertoire Version 2 (MSR2) for the Development of Label Generation Rules for the Root Zone https://www.icann.org/news/announcement220150427en

- Thai Industrial Standard (TIS) 1566 – Thai Input/ Output Methods for Computers http://www.ratchakitcha.soc.go.th/DATA/PDF/2542/E/088/9.PDF
- Computers and the Thai Language http://lexitron.nectec.or.th/KM_HL5001/file_HL5001/Paper/Inter%20Journal/krrn_52085.pdf
- Thai EGIDS http://www.thaischool1.in.th/_files_school/30104270/data/30104270_1_20140504-051017.pdf

Appendix A: Code Points Short-listed in the LGR Proposal

The code points of the Thai script shortlisted in MSR-2 are shown below. The code cells with yellow highlighting are part of the MSR, whereas those with pink highlighting are excluded from the MSR and those without highlighting are for code points that are not PVALID in IDNA 2008. A red X shows code points excluded from this LGR.



	0E0	0E1	0E2	0E3	0E4	0E5	0E6	0E7
0	▨	ฐ 0E10	ภ 0E20	ะ 0E30	เ 0E40	๐ 0E50	▨	▨
1	ก 0E01	ฑ 0E11	ม 0E21	◌ั 0E31	แ 0E41	๑ 0E51	▨	▨
2	ข 0E02	ฒ 0E12	ย 0E22	า 0E32	โ 0E42	๒ 0E52	▨	▨
3	ฃ 0E03	ณ 0E13	ร 0E23	ำ 0E33	ใ 0E43	๓ 0E53	▨	▨
4	ค 0E04	ด 0E14	ฤ 0E24	◌ิ 0E34	ไ 0E44	๔ 0E54	▨	▨
5	ฅ 0E05	ต 0E15	ล 0E25	◌ี 0E35	ๅ 0E45	๕ 0E55	▨	▨
6	ฆ 0E06	ถ 0E16	ฦ 0E26	◌ึ 0E36	☒ ๆ 0E46	๖ 0E56	▨	▨
7	ง 0E07	ท 0E17	ว 0E27	◌ื 0E37	◌็ 0E47	๗ 0E57	▨	▨
8	จ 0E08	ธ 0E18	ศ 0E28	◌ุ 0E38	◌่ 0E48	๘ 0E58	▨	▨
9	ฉ 0E09	น 0E19	ษ 0E29	◌ู 0E39	◌้ 0E49	๙ 0E59	▨	▨
A	ช 0E0A	บ 0E1A	ส 0E2A	◌ฺ 0E3A	◌๊ 0E4A	๚ 0E5A	▨	▨
B	ซ 0E0B	ป 0E1B	ห 0E2B	▨	◌๋ 0E4B	๛ 0E5B	▨	▨
C	ฌ 0E0C	ผ 0E1C	ฬ 0E2C	▨	◌์ 0E4C	▨	▨	▨
D	ญ 0E0D	ฝ 0E1D	อ 0E2D	▨	◌ํ 0E4D	▨	▨	▨
E	ฎ 0E0E	พ 0E1E	ฮ 0E2E	▨	☒ ◌๎ 0E4E	▨	▨	▨
F	ฏ 0E0F	ฟ 0E1F	ฯ 0E2F	฿ 0E3F	๏ 0E4F	▨	▨	▨

□ Excluded by IDNA 2008 or a Digit

🟥 Excluded by MSR

☒ Proposed for exclusion from LGR

🟨 Proposed for inclusion in LGR

Appendix B: Cross-Script Homoglyph Analysis Tables

Thai and Khmer homoglyphs Analysis

The visual perception of the letters listed in the following table were taken into account in the homoglyph analysis. There are some similarity in the visual perception of some of the Khmer characters and certain Thai characters or Thai letter combinations.

Khmer Character	Unicode	Thai Character	Unicode
ក	1780	ก + ◌็	0E01 + 0E47
គ	1782	ค+◌็	0E04 + 0E47
ឃ	1783	ช+ช, ฬ	0E0A+0E0A 0E2C
ង	1784	ฬ	0E2C
ឈ	1788	ถ+ช+ช	0E16+0E0A+ 0E0A
ដ	178A	ผ+◌ั	0E1C+0E31
ឋ	178B	ช ซ	0E0A 0E0B
ឍ	178D	ฌ ฒ ต+ร	0E0C 0E12 0E15+0E23
ណ	178E	ฌ+ก ญ	0E0C+0E01 0E0C
ត	178F	ด+◌๊	0E14+0E4A
ន	1793	ร	0E23
ប	1794	ช ย	0E0A 0E22
ព	1796	ด ต ถ	0E14 0E15 0E16
ភ	1797	ภ + ◌็	0E20 + 0E47
ម	1798	ษ ย	0E29 0E22
យ	1799	ผ	0E1C
រ	179A	ร	0E23
ល	179B	ญ	0E0C
វ	179C	ร	0E23
ឝ	179D	ศ + ◌็	0E28+0E47
ឞ	179E	ษ ย	0E29 0E22
ហ	17A0	ย+า	0E22+0E32

Khmer Character	Unicode	Thai Character	Unicode
អ	17A2	ร+ร	0E23+0E23
ឥ	17A5	ส	0E2A
ឫ	17AB	ช+◌ุ ย+◌ุ	0E0A+0E38 0E22+0E38
ឬ	17AC	ช+◌ุ	0E0A+0E38
ឭ	17AD	ฅ+◌ุ ต+◌ุ	0E05+0E38 0E15+0E38
ឮ	17AE	ฅ+◌ุ ต+◌ุ	0E05+0E38 0E15+0E38
ឯ	17AF	ฉ ฬ	0E09 0E2C
◌ា	17B6	า ๅ	0E32 0E45
◌ិ	17B7	◌ิ	0E34
◌ី	17B8	◌ี	0E35
◌ឹ	17B9	◌ึ	0E36
◌ឺ	17BA	◌ื	0E37
◌ុ	17BB	◌ุ	0E38
◌ូ	17BC	◌ู	0E39
◌ួ	17BD	◌ู	0E39
◌ើ	17BE	เ+◌ี	0E40+0E35
◌េ	17C1	เ	0E40
◌ោ	17C4	เ+า เ+ๅ	0E40+0E32 0E40+0E45
◌ៅ	17C5	เ+า เ+ๅ	0E40+0E32 0E40+0E45
◌េ	17C1	เ	0E40
◌ោ	17C4	เ+า เ+ๅ	0E40+0E32 0E40+0E45
◌ៅ	17C5	เ+า เ+ๅ	0E40+0E32 0E40+0E45

Khmer Character	Unicode	Thai Character	Unicode
◌ំ	17C6	◌ํ	0E4D
◌ះ	17C7	ะ	0E30
◌ៈ	17C8	ะ	0E30
◌៊	17CA	◌๊	0E4A
◌់	17CB	◌่	0E48

Khmer Character	Unicode	Thai Character	Unicode
◌៌	17CC	◌็	0E47
◌៍	17CD	◌์	0E4C
◌៎	17CE	◌๋	0E4B
◌៏	17CF	◌็	0E47
◌័	17D0	◌ั	0E31

Thai and Myanmar homoglyphs Analysis

The visual perception of the letters listed in the following table were taken into account in the homoglyph analysis. There is some similarity in visual perception between some of the Myanmar characters and certain Thai character or Thai letter combinations. The detailed information is listed on table below

Myanmar Character	Unicode	Thai Character	Unicode
ခ	1001	ว อ	0E27 0E2D
ဂ	1002	ก	0E01
ဃ	1003	พ ฟ	0E1E 0E1F
ဈ	1008	ข ฃ	0E02 0E03
ဎ	100E	ข ฃ บ	0E02 0E03 0E1A
ပ	1015	ข บ	0E02 0E1A
ဘ	1018	ว + ว	0E27 + 0E27
ယ	101A	ผ พ	0E1C 0E1E
ရ	101B	ฤ ใ	0E24 0E43
ါ	102B	า ๅ	0E32 0E45
◌ိ	102D	◌ํ	0E4D
◌ံ	1036	◌ํ	0E4D
◌့	1037	◌ฺ	0E3A
း	1038	ะ	0E30

Myanmar Character	Unicode	Thai Character	Unicode
၅	1045	ๆ	0E46
၈	1048	ด	0E14
၊	104A	เ	0E40
။	104B	แ	0E41
◌ၢ	1062	า	0E32
ၤ	1064	า	0E32
ၫ	106B	า	0E32
◌ၵ	1075	ภ	0E20
◌ၶ	1076	ว	0E27
◌ၷ	1077	ภ	0E20
ႀ	1080	ม	0E21
ႃ	1083	ๅ	0E45
◌ႊ	108A	ะ	0E30

Thai and Lao homoglyphs Analysis

The visual perception of the letters listed in the following table were taken into account in the homoglyph analysis. There is some similarity in the visual perception of between some of the Lao Characters and certain Thai Characters or Thai letter combinations. The detailed information is listed below

Lao Character	Unicode	Thai Character	Unicode
ກ	0E81	ท	0E17
ຄ	0E84	ถ	0E16
ຈ	0E88	จ	0E08
ຍ	0E8D	ย	0E22
ດ	0E94	ถ	0E16
ຕ	0E95	ต	0E15
ຖ	0E96	ถ ฤ	0E16 0E24
ທ	0E97	ท	0E17
ນ	0E99	ม	0E21
ບ	0E9A	บ	0E1A
ປ	0E9B	ป	0E1B
ຜ	0E9C	ผ	0E1C
ຝ	0E9D	ฝ	0E1D
ພ	0E9E	พ	0E1E
ຟ	0E9F	ฟ	0E1F
ມ	0EA1	ม	0E21
ຢ	0EA2	ย	0E22
ຣ	0EA3	ธ ร	0E18 0E23
ລ	0EA5	ล	0E25
ວ	0EA7	ว อ	0E27 0E2D
ສ	0EAA	ส	0E2A

Lao Character	Unicode	Thai Character	Unicode
ຫ	0EAB	ท ห	0E17 0E2B
ອ	0EAD	ฮ	0E2E
ຮ	0EAE	ธ ร	0E18 0E23
ະ	0EB0	ะ	0E30
ັ	0EB1	◌ั	0E31
າ	0EB2	า	0E32
ຳ	0EB3	◌ำ	0E33
ຸ	0EB8	◌ุ	0E38
ູ	0EB9	◌ู	0E39
ົ	0EBB	◌์	0E4C
ເ	0EC0	เ	0E40
ແ	0EC1	แ	0E41
ໂ	0EC2	โ	0E42
ໃ	0EC3	ใ	0E43
ໄ	0EC4	ไ	0E44
່	0EC8	◌่	0E48
້	0EC9	◌้	0E49
໊	0ECA	◌๊	0E4A
໋	0ECB	◌๋	0E4B
໌	0ECC	◌์	0E4C
ໍ	0ECD	◌ํ	