

ENS Name Normalization

raffy.eth / @adraffy



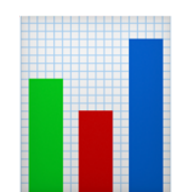
Since Last Update — Oct 2022



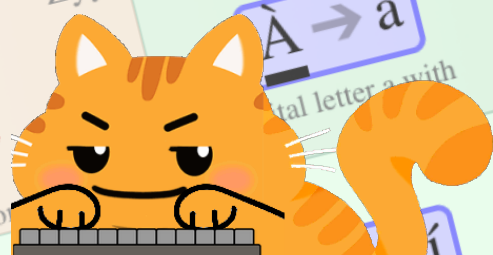
ENSIP-15



Future Stuff



Status Report



- **Added: Confusable Implementation**
- **Changed: Combining Marks → Nonspacing Marks**
-  **Spec Finalized @ `ens-normalize.js v1.9.0`**
 - Unmodified since **Feb 25, 2023**
 - SHA-256: 962316964553fce6188e25a5166a4c1e906333adf53bdf2964c71dedc0f8e2c8
- **Updated: Documentation**
- **Deployed in `ethers`, `viem`, and `app.ens.domains`!**



Since Last Update

Confusables

<https://adraffy.github.io/ens-normalize.js/test/confused.html>

✗ Normalized: 0x ⚠ whole-script confusable: Cyrillic/Latin

Codepoint	Form	Character Name	Script	Latin	Han	Japanese	Korean	Cyrillic	Greek	Cans	Runr	Ital	Tfng	Lisu	Carl	Elba	Lyci	Sarb	Wara
30	0	DIGIT ZERO	Common	A	A	A	A	A	A										
78	X	LATIN SMALL LETTER X	Latin	A	A	A	A												
445	X	CYRILLIC SMALL LETTER HA	Cyrillic					B											
4B3	X	CYRILLIC SMALL LETTER HA WITH DESCENDER	Cyrillic					B											
1541	x	CANADIAN SYLLABICS SAYISI YI	Canadian Aboriginal							C									
157D	x	CANADIAN SYLLABICS HK	Canadian Aboriginal							C									
166D	X	CANADIAN SYLLABICS CHI SIGN	Canadian Aboriginal							C									
16B7	X	RUNIC LETTER GEBO GYFU G	Runic								D								
10322	X	OLD ITALIC NUMERAL TEN	Old Italic									E							
10317	X	OLD ITALIC LETTER EKS	Old Italic									E							
2D5D	X	TIFINAGH LETTER YATH	Tifinagh										F						
A4EB	X	LISU LETTER SHA	Lisu											G					
102B4	X	CARIAN LETTER X	Carian												H				
10527	X	ELBASAN LETTER KHE	Elbasan													I			
10290	X	LYCIAN LETTER MM	Lycian														J		
10A69	X	OLD SOUTH ARABIAN LETTER TAW	Old South Arabian															K	
118EC	X	WARANG CITI NUMBER THIRTY	Warang Citi																L

Latin	Han	Japanese	Korean	Cyrillic
A	A	A	A	A
A	A	A	A	
				B
				B

0x 2ch 30 445

However...

0x ASCII 2ch 30 78

🔧 Since Last Update

Confusables

<https://adraffy.github.io/ens-normalize.js/test/confused.html>

✓ Normalized: **тæ**

тæ
Cyrillic

2_{ch} 442 4D5

Codepoint	Form	Character Name	Script	Latin	Cyrillic	Greek
442	Т	CYRILLIC SMALL LETTER TE	Cyrillic		A	
3C4	τ	GREEK SMALL LETTER TAU	Greek			B
E6	æ	LATIN SMALL LETTER AE	Latin	A		
4D5	æ	CYRILLIC SMALL LIGATURE A IE	Cyrillic		B	

Latin	Greek
×	B
A	×

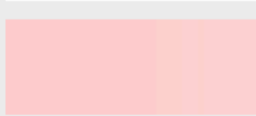
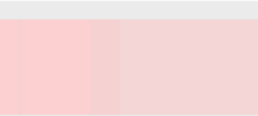
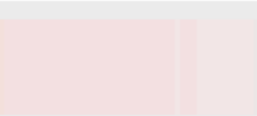
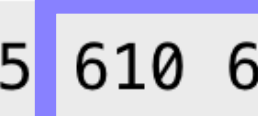
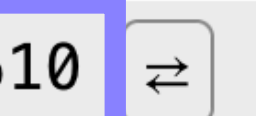
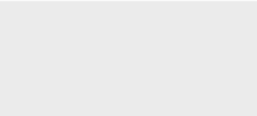
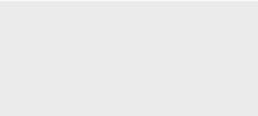
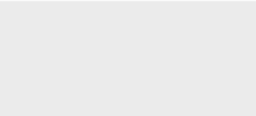
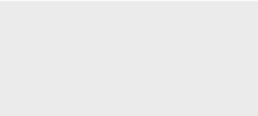
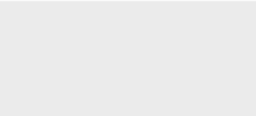
Since Last Update

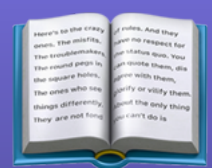
CM → **NSM**

- **NSM** are a subset of **CM** that doesn't increase character "width"
- **Must be Unique** and **Limit 4** (Unicode UTS-39)

Input:   eth  ENSIP-1:  Diff Norm  Invalid label "ٲ": non-spacing marks: repeated "ٲ" {610}

Namehash: 0xd32ee6f09121d7e7688e0634581702d3474953b1dac9e1771d5ebacdca327ab8



<https://github.com/adraffy/ensip-15>

- Implementation Details
- Derivation of Spec
- Description of `spec.json`
- Provides many examples
- Links to all Unicode resources

- Unicode version `15.0.0`
- `spec.json` contains all `necessary data` for normalization.

- **Groups** (`"groups"`) — `groups` of characters that can constitute a label
 - `"name"` — ASCII name of the group (or abbreviation if **Restricted**)
 - Examples: *Latin*, *Japanese*, *Egyp*
 - **Restricted** (`"restricted"`) — `true` if **Excluded** or **Limited-Use** script
 - Examples: *Latin* → `false`, *Egyp* → `true`
 - `"primary"` — subset of characters that define the group
 - Examples: `"a"` → *Latin*, `"あ"` → *Japanese*, `"𐤀"` → *Egyp*
 - `"secondary"` — subset of characters included with the group
 - Example: `"0"` → *Common* but mixable with *Latin*
 - **CM Whitelist(ed)** (`"cm"`) — (optional) set of allowed compound sequences in NFC
 - Each compound sequence is a character followed by one or more **Combining Marks**.
 - Example: `â` → `E0 300 300`
 - Currently, every group that is **CM Whitelist** has zero compound sequences.
 - **CM Whitelisted** is effectively `true` if `[]` otherwise `false`
- **Ignored** (`"ignored"`) — characters that are ignored during normalization
 - Example: `34F (0) COMBINING GRAPHEME JOINER`
- **Mapped** (`"mapped"`) — characters that are mapped to a sequence of **valid** characters
 - Example: `41 (A) LATIN CAPITAL LETTER A` → `[61 (a) LATIN SMALL LETTER A]`
 - Example: `2165 (VI) ROMAN NUMERAL SIX` → `[76 (v) LATIN SMALL LETTER V, 69 (i) LATIN SMALL LETTER I]`
- **Whole Confusable** (`"wholes"`) — groups of characters that look similar
 - `"valid"` — subset of confusable characters that are allowed
 - Example: `34 (4) DIGIT FOUR`
 - **Confused** (`"confused"`) — subset of confusable characters that confuse
 - Example: `13CE (4) CHEROKEE LETTER SE`
- **Fenced** (`"fenced"`) — set of characters that cannot be first, last, or contiguous
 - Example: `2044 (/) FRACTION SLASH`
- **Emoji Sequences** (`"emoji"`) — `allowed` emoji sequences
 - Example: `👨 [1F468 200D 1F4BB] man technologist`
- **Combining Marks / CM** (`"cm"`) — characters that are **Combining Marks**
- **Non-spacing Marks / NSM** (`"nsm"`) — valid subset of **CM** with general category (`"Mn"` or `"Me"`)
- **Maximum NSM** (`"nsm_max"`) — maximum sequence length of unique **NSM**
- **Should Escape** (`"escape"`) — characters that shouldn't be printed
- **NFC Check** (`"nfc_check"`) — valid characters that `may require NFC`



ENSIP-15

Documentation

- IDNA 2003

- UseSTD3ASCIIRules is **true**
- VerifyDnsLength is **false**
- Transitional_Processing is **false**
- The following **deviations** are **valid**:
 - DF (ß) LATIN SMALL LETTER SHARP S
 - 3C2 (ς) GREEK SMALL LETTER FINAL SIGMA
- CheckHyphens is **false** ([WhatWG URL Spec § 3.3](#))
- CheckBidi is **false**
- ContextJ:
 - 200C ( ) ZERO WIDTH NON-JOINER (ZWNJ) is **disallowed everywhere**.
 - 200D (⁠) ZERO WIDTH JOINER (ZWJ) is **only allowed** in emoji sequences.
- ContextO:
 - B7 (·) MIDDLE DOT is **disallowed**.
 - 375 (,) GREEK LOWER NUMERAL SIGN is **disallowed**.
 - 5F3 (׳) HEBREW PUNCTUATION GERESH and 5F4 (״) HEBREW PUNCTUATION GERSHAYIM are *Greek*.
 - 30FB (・) KATAKANA MIDDLE DOT is **Fenced** and *Han, Japanese, Korean, and Bopomofo*.
 - Some **Extended Arabic Numerals** are **mapped**:
 - 6F0 (٠) → 660 (٠) ARABIC-INDIC DIGIT ZERO
 - 6F1 (١) → 661 (١) ARABIC-INDIC DIGIT ONE
 - 6F2 (٢) → 662 (٢) ARABIC-INDIC DIGIT TWO
 - 6F3 (٣) → 663 (٣) ARABIC-INDIC DIGIT THREE
 - 6F7 (٧) → 667 (٧) ARABIC-INDIC DIGIT SEVEN
 - 6F8 (٨) → 668 (٨) ARABIC-INDIC DIGIT EIGHT
 - 6F9 (٩) → 669 (٩) ARABIC-INDIC DIGIT NINE

- **Punycode** is not decoded.

- **Many characters** are **disallowed** for various reasons:

- Nearly all punctuation are **disallowed**.
 - Example: 589 (:) ARMENIAN FULL STOP
- All parentheses and brackets are **disallowed**.
 - Example: 2997 (()) LEFT BLACK TORTOISE SHELL BRACKET
- Nearly all vocalization annotations are **disallowed**.
 - Example: 294 (ʔ) LATIN LETTER GLOTTAL STOP
- Obsolete, deprecated, and ancient characters are **disallowed**.
 - Example: 463 (ѣ) CYRILLIC SMALL LETTER YAT
- Combining, modifying, reversed, flipped, turned, and partial variations are **disallowed**.
 - Example: 218A (𐍺) TURNED DIGIT TWO
- When multiple weights of the same character exist, the variant closest to "heavy" is selected.
 - Example: 271A (⚡) HEAVY GREEK CROSS
 - This occasionally selects an emoji.
 - Example: ✓ or 2714 (✓) HEAVY CHECK MARK is selected instead of 2715 (✓) CHECK MARK
- Many visually confusable characters are **disallowed**.
 - Example: 131 (ı) LATIN SMALL LETTER DOTLESS I
- Many ligatures, *n*-graphs, and *n*-grams are **disallowed**.
 - Example: A74F (∞) LATIN SMALL LETTER OO
- Many esoteric characters are **disallowed**.
 - Example: 2376 (Ɑ) APL FUNCTIONAL SYMBOL ALPHA UNDERBAR



Since Last Update

Extra Docs

Additional Disallowed Characters (5877)

Codepoint	Form	Name
200C		ZERO WIDTH NON-JOINER
200D		ZERO WIDTH JOINER
3002	。	IDEOGRAPHIC FULL STOP
FF0E	.	FULLWIDTH FULL STOP
FF61	。	HALFWIDTH IDEOGRAPHIC FULL STOP
589	:	ARMENIAN FULL STOP
61D	□	ARABIC END OF TEXT MARK
61E	⋯	ARABIC TRIPLE DOT PUNCTUATION MARK
6D4	.	ARABIC FULL STOP
700	⋮	SYRIAC END OF PARAGRAPH
701	⋯	SYRIAC SUPRALINEAR FULL STOP
702	⋯	SYRIAC SUBLINEAR FULL STOP
964	।	DEVANAGARI DANDA
1362	::	ETHIOPIC FULL STOP
166E	x	CANADIAN SYLLABICS FULL STOP
1803	⋯	MONGOLIAN FULL STOP
1809	»	MONGOLIAN MANCHU FULL STOP
2CF9	Ⲙ	COPTIC OLD NUBIAN FULL STOP
2CFE	Ⲙ	COPTIC FULL STOP
2E3C	□	STENOGRAPHIC FULL STOP
A4FF	=	LISU PUNCTUATION FULL STOP
A60E	*	VAI FULL STOP
A6F3	𑌆	BAMUM FULL STOP
16AF5	+	BASSA VAH FULL STOP
16F00	𑌀	MEDEFAIDRI FULL STOP

Supported Emoji

- 1. [Valid Emoji Sequences \(3637\)](#)
- 2. [Disabled Emoji Characters \(33\)](#)
- 3. [Demoted Emoji Characters \(32\)](#)
- 4. [Emoji Sequence Whitelist \(0\)](#)
- 5. [Emoji Sequence Blacklist \(0\)](#)

Valid Emoji Sequences (3637)

Codepoints	Form	Name
A9 FE0F	©	copyright
AE FE0F	®	registered
2194 FE0F	↔	left-right arrow
2195 FE0F	↕	up-down arrow
2196 FE0F	↖	up-left arrow
2197 FE0F	↗	up-right arrow
2198 FE0F	↘	down-right arrow
2199 FE0F	↙	down-left arrow
21A9 FE0F	↪	right arrow curving left
21AA FE0F	↩	left arrow curving right
231A FE0F	🕒	watch
231B FE0F	🕒	hourglass done
2328 FE0F	🖨	keyboard
23CF FE0F	🗑	eject button
23E9 FE0F	⏮	fast-forward button
23EA FE0F	⏮	fast reverse button
23F0 FE0F	⏮	fast reverse button

Groups (28)

Group Name	Primary Scripts	Secondary Scripts	Restricted CM	Extra Characters
Latin	Latin, Common	Inherited	✓	π
Arabic	Arabic			–
Han	Han	Common	✓	abcdefghijklmnopqrstuvwxyz
Devanagari	Devanagari			
Japanese	Katakana, Hiragana	Han, Common	✓	abcdefghijklmnopqrstuvwxyz
Korean	Hangul	Han, Common	✓	abcdefghijklmnopqrstuvwxyz
Cyrillic	Cyrillic	Common, Inherited	✓	
Hebrew	Hebrew			ℵ
Thai	Thai			฿
Bengali	Bengali			
Tamil	Tamil			
Greek	Greek	Common, Inherited	✓	
Tibetan	Tibetan			
Gurmukhi	Gurmukhi			
Lao	Lao			
Gujarati	Gujarati			
Telugu	Telugu			
Bopomofo	Bopomofo	Han	✓	
Myanmar	Myanmar			
Georgian	Georgian			
Armenian	Armenian			
Malayalam	Malayalam			
Khmer	Khmer			
Kannada	Kannada			



Future Stuff

Wide Confusables

- This ENSIP only addresses **single-character** **confusables**.

- There exist confusable **multi-character** sequences:

- "𐄂" [BB6 BCD BB0 BC0]

- "𐄂" [BB8 BCD BB0 BC0]

- There exist confusable emoji sequences:

- 🚲 [1F6B4] and 🚲 [1F6B4 1F3FB]

- 🇺🇸 [1F1FA 1F1F8] and 🇺🇸 [1F1FA 1F1F2]

- ♥ [2665] BLACK HEART SUIT and ♥ [2764] HEAVY BLACK HEART

Unicode 15.1/16 Emoji Direction

Internal Representation			Intended Display	Fallback Appearance
	ZW J			 
U+1F3C3	U+200D	U+2B05 U+FE0F		 
	ZW J			
U+1F3C3	U+200D	U+27A1 U+FE0F		

Fallback appearance is
indistinguishable
from **two** separate emoji.

Bidi Label Ordering (Rare)

Consider the following labels:

1. Directionless: "123" (Bidi class = EN)
2. RTL: "سلمان" (Bidi class = AL)
3. LTR: "eth" (Bidi class = L)

Consider the following names:

- [1.2.3] "123.سلمان.eth"

◦ Parsed: 123.سلمان.eth

◦ Display: 123.سلمان.eth

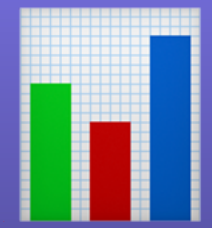
- [2.1.3] "123.سلمان.eth"

◦ Parsed: سلمان.123.eth

◦ Display: 123.سلمان.eth

Different

Same



Status Report

Summary



Implementation

<https://github.com/adraffy/ens-normalize.js>



Documentation: ENSIP-15

<https://github.com/adraffy/ensip-15>



Invalid Name Refund Proposal by [matoken.eth](#)

<https://discuss.ens.domains/t/temp-check-ens-invalid-name-refund-proposal/16824>



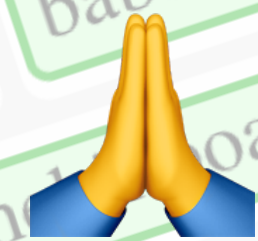
Draft Proposal for Social Vote

<https://discuss.ens.domains/t/draft-approval-of-ens-name-normalization-standard/16957>

2023-06-01

ENS Name Normalization

raffy.eth / @adraffy



Thanks for listening!

