# HTML Encoding

Ing. Noé Rodríguez Castro
nrodriguezc@up.edu.mx
@noahrod

# Historia

- ASCII fue el primer estándar de codificación de caracteres.

- **¿Por qué ya no se usa?**
  - Porque no todo el mundo habla inglés. (Thank God!)

| Dec | Hx | Oct | Char | Dec | Hx | Oct | Html | Chr | Dec | Hx | Oct | Html | Chr | Dec | Hx | Oct | Html | Chr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 000 | NUL (null) | 32 | 20 | 040 |   | Space | 64 | 40 | 100 | @ | @ | 96 | 60 | 140 | ` | ` |
| 1 | 1 | 001 | SOH (start of heading) | 33 | 21 | 041 | ! | ! | 65 | 41 | 101 | A | A | 97 | 61 | 141 | a | a |
| 2 | 2 | 002 | STX (start of text) | 34 | 22 | 042 | " | " | 66 | 42 | 102 | B | B | 98 | 62 | 142 | b | b |
| 3 | 3 | 003 | ETX (end of text) | 35 | 23 | 043 | # | # | 67 | 43 | 103 | C | C | 99 | 63 | 143 | c | c |
| 4 | 4 | 004 | EOT (end of transmission) | 36 | 24 | 044 | $ | $ | 68 | 44 | 104 | D | D | 100 | 64 | 144 | d | d |
| 5 | 5 | 005 | ENQ (enquiry) | 37 | 25 | 045 | % | % | 69 | 45 | 105 | E | E | 101 | 65 | 145 | e | e |
| 6 | 6 | 006 | ACK (acknowledge) | 38 | 26 | 046 | & | & | 70 | 46 | 106 | F | F | 102 | 66 | 146 | f | f |
| 7 | 7 | 007 | BEL (bell) | 39 | 27 | 047 | ' | ' | 71 | 47 | 107 | G | G | 103 | 67 | 147 | g | g |
| 8 | 8 | 010 | BS (backspace) | 40 | 28 | 050 | ( | ( | 72 | 48 | 110 | H | H | 104 | 68 | 150 | h | h |
| 9 | 9 | 011 | TAB (horizontal tab) | 41 | 29 | 051 | ) | ) | 73 | 49 | 111 | I | I | 105 | 69 | 151 | i | i |
| 10 | A | 012 | LF (NL line feed, new line) | 42 | 2A | 052 | * | * | 74 | 4A | 112 | J | J | 106 | 6A | 152 | j | j |
| 11 | B | 013 | VT (vertical tab) | 43 | 2B | 053 | + | + | 75 | 4B | 113 | K | K | 107 | 6B | 153 | k | k |
| 12 | C | 014 | FF (NP form feed, new page) | 44 | 2C | 054 | , | , | 76 | 4C | 114 | L | L | 108 | 6C | 154 | l | l |
| 13 | D | 015 | CR (carriage return) | 45 | 2D | 055 | - | - | 77 | 4D | 115 | M | M | 109 | 6D | 155 | m | m |
| 14 | E | 016 | SO (shift out) | 46 | 2E | 056 | . | . | 78 | 4E | 116 | N | N | 110 | 6E | 156 | n | n |
| 15 | F | 017 | SI (shift in) | 47 | 2F | 057 | / | / | 79 | 4F | 117 | O | O | 111 | 6F | 157 | o | o |
| 16 | 10 | 020 | DLE (data link escape) | 48 | 30 | 060 | 0 | 0 | 80 | 50 | 120 | P | P | 112 | 70 | 160 | p | p |
| 17 | 11 | 021 | DC1 (device control 1) | 49 | 31 | 061 | 1 | 1 | 81 | 51 | 121 | Q | Q | 113 | 71 | 161 | q | q |
| 18 | 12 | 022 | DC2 (device control 2) | 50 | 32 | 062 | 2 | 2 | 82 | 52 | 122 | R | R | 114 | 72 | 162 | r | r |
| 19 | 13 | 023 | DC3 (device control 3) | 51 | 33 | 063 | 3 | 3 | 83 | 53 | 123 | S | S | 115 | 73 | 163 | s | s |
| 20 | 14 | 024 | DC4 (device control 4) | 52 | 34 | 064 | 4 | 4 | 84 | 54 | 124 | T | T | 116 | 74 | 164 | t | t |
| 21 | 15 | 025 | NAK (negative acknowledge) | 53 | 35 | 065 | 5 | 5 | 85 | 55 | 125 | U | U | 117 | 75 | 165 | u | u |
| 22 | 16 | 026 | SYN (synchronous idle) | 54 | 36 | 066 | 6 | 6 | 86 | 56 | 126 | V | V | 118 | 76 | 166 | v | v |
| 23 | 17 | 027 | ETB (end of trans. block) | 55 | 37 | 067 | 7 | 7 | 87 | 57 | 127 | W | W | 119 | 77 | 167 | w | w |
| 24 | 18 | 030 | CAN (cancel) | 56 | 38 | 070 | 8 | 8 | 88 | 58 | 130 | X | X | 120 | 78 | 170 | x | x |
| 25 | 19 | 031 | EM (end of medium) | 57 | 39 | 071 | 9 | 9 | 89 | 59 | 131 | Y | Y | 121 | 79 | 171 | y | y |
| 26 | 1A | 032 | SUB (substitute) | 58 | 3A | 072 | : | : | 90 | 5A | 132 | Z | Z | 122 | 7A | 172 | z | z |
| 27 | 1B | 033 | ESC (escape) | 59 | 3B | 073 | ; | ; | 91 | 5B | 133 | [ | [ | 123 | 7B | 173 | { | { |
| 28 | 1C | 034 | FS (file separator) | 60 | 3C | 074 | < | < | 92 | 5C | 134 | \ | \ | 124 | 7C | 174 | | | \| |
| 29 | 1D | 035 | GS (group separator) | 61 | 3D | 075 | = | = | 93 | 5D | 135 | ] | ] | 125 | 7D | 175 | } | } |
| 30 | 1E | 036 | RS (record separator) | 62 | 3E | 076 | > | > | 94 | 5E | 136 | ^ | ^ | 126 | 7E | 176 | ~ | ~ |
| 31 | 1F | 037 | US (unit separator) | 63 | 3F | 077 | ? | ? | 95 | 5F | 137 | _ | _ | 127 | 7F | 177 | &#127; | DEL |

Source: www.LookupTables.com

# Historia II

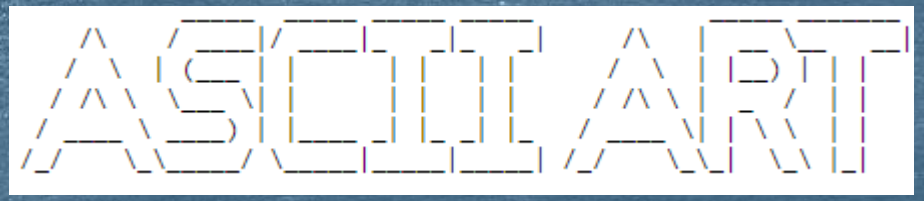
- Obviamente intentaron solucionarlo…. (Extended ASCII)

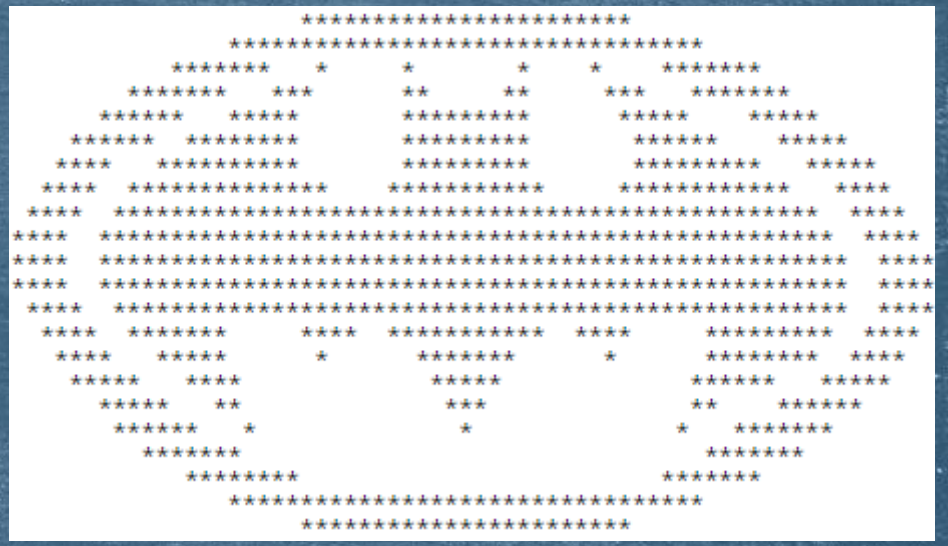
| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 128 | Ç | 144 | É | 160 | á | 176 | ▒ | 192 | └ | 208 | ╨ | 224 | α | 240 | ≡ |
| 129 | ü | 145 | æ | 161 | í | 177 | ▒ | 193 | ┴ | 209 | ╤ | 225 | ß | 241 | ± |
| 130 | é | 146 | Æ | 162 | ó | 178 | ▓ | 194 | ┬ | 210 | ╥ | 226 | Γ | 242 | ≥ |
| 131 | â | 147 | ô | 163 | ú | 179 | │ | 195 | ├ | 211 | ╙ | 227 | π | 243 | ≤ |
| 132 | ä | 148 | ö | 164 | ñ | 180 | ┤ | 196 | ─ | 212 | ╘ | 228 | Σ | 244 | ⌠ |
| 133 | à | 149 | ò | 165 | Ñ | 181 | ╡ | 197 | ┼ | 213 | ╒ | 229 | σ | 245 | ⌡ |
| 134 | å | 150 | û | 166 | ª | 182 | ╢ | 198 | ╞ | 214 | ╓ | 230 | µ | 246 | ÷ |
| 135 | ç | 151 | ù | 167 | º | 183 | ╖ | 199 | ╟ | 215 | ╫ | 231 | τ | 247 | ≈ |
| 136 | ê | 152 | ÿ | 168 | ¿ | 184 | ╕ | 200 | ╚ | 216 | ╪ | 232 | Φ | 248 | ° |
| 137 | ë | 153 | Ö | 169 | ⌐ | 185 | ╣ | 201 | ╔ | 217 | ┘ | 233 | Θ | 249 | ∙ |
| 138 | è | 154 | Ü | 170 | ¬ | 186 | ║ | 202 | ╩ | 218 | ┌ | 234 | Ω | 250 | · |
| 139 | ï | 155 | ¢ | 171 | ½ | 187 | ╗ | 203 | ╦ | 219 | █ | 235 | δ | 251 | √ |
| 140 | î | 156 | £ | 172 | ¼ | 188 | ╝ | 204 | ╠ | 220 | ▄ | 236 | ∞ | 252 | ⁿ |
| 141 | ì | 157 | ¥ | 173 | ¡ | 189 | ╜ | 205 | ═ | 221 | ▌ | 237 | φ | 253 | ² |
| 142 | Ä | 158 | ₧ | 174 | « | 190 | ╛ | 206 | ╬ | 222 | ▐ | 238 | ε | 254 | ■ |
| 143 | Å | 159 | ƒ | 175 | » | 191 | ┐ | 207 | ╧ | 223 | ▀ | 239 | ∩ | 255 | |

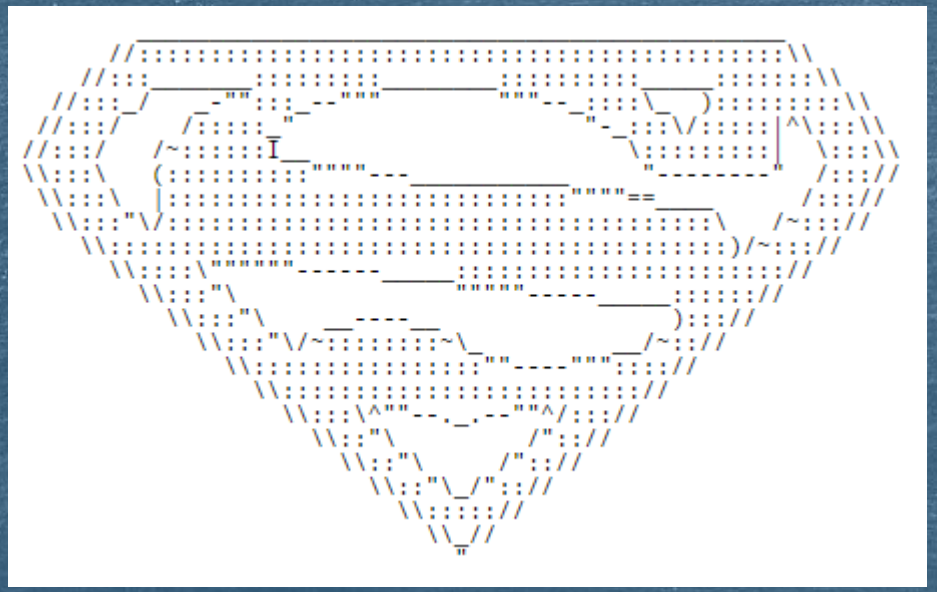
Source: www.LookupTables.com

# Historia III

- De hecho hoy en día se usa para hacer “arte”.

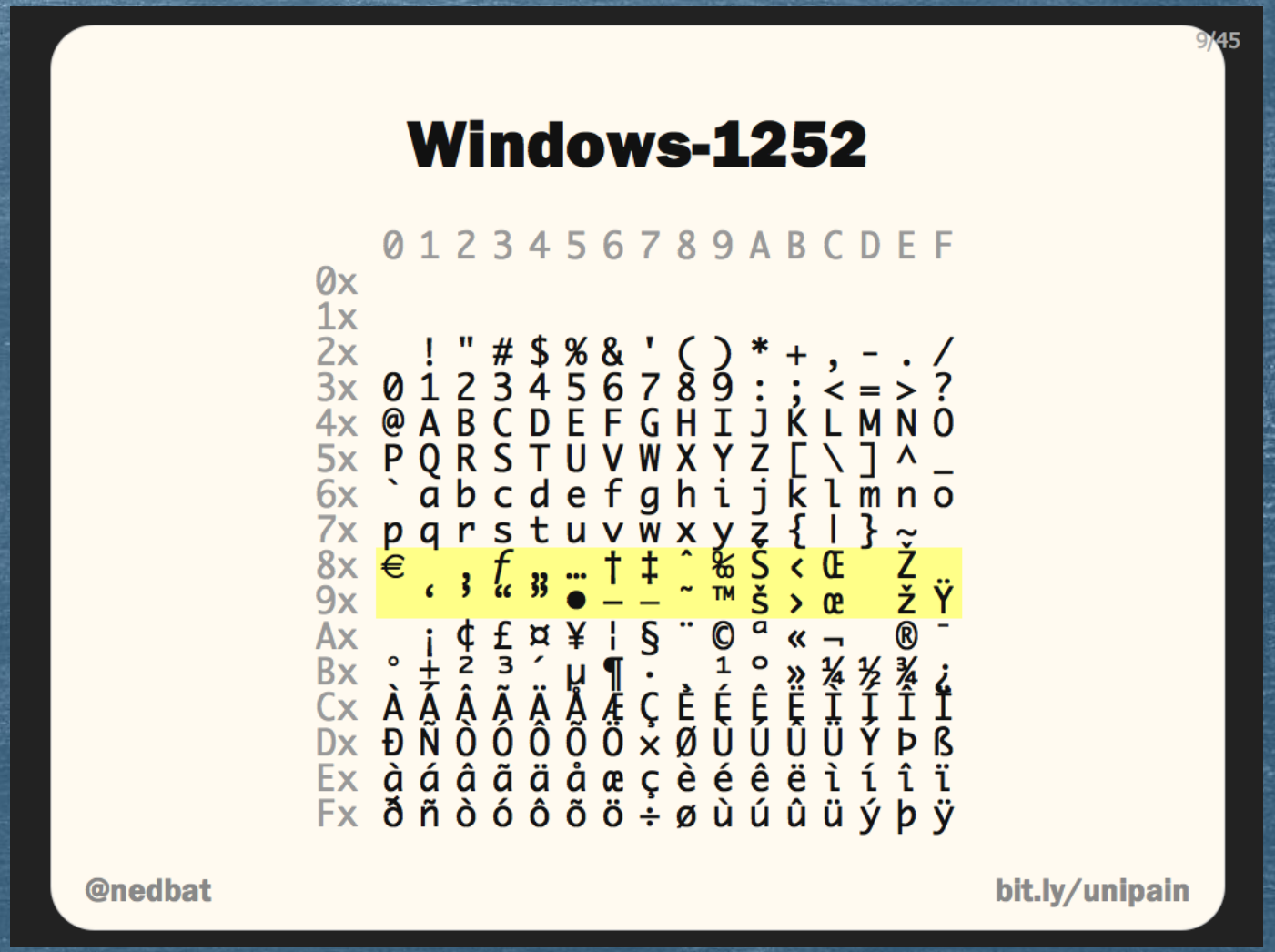




# Historia IV

- Windows con sus ventas fue de los primeros en darse cuenta que necesitaba un charset mas completo para satisfacer a sus clientes alrededor de mundo y nos presento la codificación “Windows-1252” que suele conocerse como “ANSI” aunque es incorrecto llamarlo así.



# Historia V

- El tema es que internet es global. Por lo que HTML4 nos presentó en enconding con aún mas chars, el "ISO-8859-1".

| Char | Code | Name | Description |
|---|---|---|---|
| | 32 | - | Normal space |
| ! | 33 | - | Exclamation |
| " | 34 | quot | Double quote |
| # | 35 | - | Hash or pound |
| $ | 36 | - | Dollar |
| % | 37 | - | Percent |
| & | 38 | - | Ampersand |
| ' | 39 | - | Apostrophe |
| ( | 40 | - | Open bracket |
| ) | 41 | - | Close bracket |
| * | 42 | - | Asterisk |
| + | 43 | - | Plus sign |
| , | 44 | - | Comma |
| - | 45 | - | Minus sign |
| . | 46 | - | Period |
| / | 47 | - | Forward slash |

| Char | Code | Name | Description |
|---|---|---|---|
| 0 | 48 | - | Digit 0 |
| 1 | 49 | - | Digit 1 |
| 2 | 50 | - | Digit 2 |
| 3 | 51 | - | Digit 3 |
| 4 | 52 | - | Digit 4 |
| 5 | 53 | - | Digit 5 |
| 6 | 54 | - | Digit 6 |
| 7 | 55 | - | Digit 7 |
| 8 | 56 | - | Digit 8 |
| 9 | 57 | - | Digit 9 |
| : | 58 | - | Colon |
| ; | 59 | - | Semicolon |
| < | 60 | lt | Less than |
| = | 61 | - | Equals |
| > | 62 | gt | Greater than |
| ? | 63 | - | Question mark |

| Char | Code | Name | Description |
|---|---|---|---|
| @ | 64 | - | At sign |
| A | 65 | - | A |
| B | 66 | - | B |
| C | 67 | - | C |
| D | 68 | - | D |
| E | 69 | - | E |
| F | 70 | - | F |
| G | 71 | - | G |
| H | 72 | - | H |
| I | 73 | - | I |
| J | 74 | - | J |
| K | 75 | - | K |
| L | 76 | - | L |
| M | 77 | - | M |
| N | 78 | - | N |
| O | 79 | - | O |

| Char | Code | Name | Description |
|---|---|---|---|
| P | 80 | - | P |
| Q | 81 | - | Q |
| R | 82 | - | R |
| S | 83 | - | S |
| T | 84 | - | T |
| U | 85 | - | U |
| V | 86 | - | V |
| W | 87 | - | W |
| X | 88 | - | X |
| Y | 89 | - | Y |
| Z | 90 | - | Z |
| [ | 91 | - | Open square bracket |
| \ | 92 | - | Backslash |
| ] | 93 | - | Close square bracket |
| ^ | 94 | - | Pointer |
| _ | 95 | - | Underscore |

| Char | Code | Name | Description |
|---|---|---|---|
| ‘ | 96 | - | Grave accent |
| a | 97 | - | a |
| b | 98 | - | b |
| c | 99 | - | c |
| d | 100 | - | d |
| e | 101 | - | e |
| f | 102 | - | f |
| g | 103 | - | g |
| h | 104 | - | h |
| i | 105 | - | i |
| j | 106 | - | j |
| k | 107 | - | k |
| l | 108 | - | l |
| m | 109 | - | m |
| n | 110 | - | n |
| o | 111 | - | o |

| Char | Code | Name | Description |
|---|---|---|---|
| p | 112 | - | p |
| q | 113 | - | q |
| r | 114 | - | r |
| s | 115 | - | s |
| t | 116 | - | t |
| u | 117 | - | u |
| v | 118 | - | v |
| w | 119 | - | w |
| x | 120 | - | x |
| y | 121 | - | y |
| z | 122 | - | z |
| { | 123 | - | Left brace |
| \| | 124 | - | Vertical bar |
| } | 125 | - | Right brace |
| ~ | 126 | - | Tilde |
| | 127 | - | (Unused) |

| Char | Code | Name | Description |
|---|---|---|---|
| | 160 | nbsp | Non-breaking space |
| ¡ | 161 | iexcl | Inverted exclamation |
| ¢ | 162 | cent | Cent sign |
| £ | 163 | pound | Pound sign |
| ¤ | 164 | curren | Currency sign |
| ¥ | 165 | yen | Yen sign |
| ¦ | 166 | brvbar | Broken bar |
| § | 167 | sect | Section sign |
| ¨ | 168 | uml | Umlaut or diaeresis |
| © | 169 | copy | Copyright sign |
| ª | 170 | ordf | Feminine ordinal |
| « | 171 | laquo | Left angle quotes |
| ¬ | 172 | not | Logical not sign |
| - | 173 | shy | Soft hyphen |
| ® | 174 | reg | Registered trademark |
| ¯ | 175 | macr | Spacing macron |

| Char | Code | Name | Description |
|---|---|---|---|
| ° | 176 | deg | Degree sign |
| ± | 177 | plusmn | Plus-minus sign |
| ² | 178 | sup2 | Superscript 2 |
| ³ | 179 | sup3 | Superscript 3 |
| ´ | 180 | acute | Spacing acute |
| µ | 181 | micro | Micro sign |
| ¶ | 182 | para | Paragraph sign |
| · | 183 | middot | Middle dot |
| ¸ | 184 | cedil | Spacing cedilla |
| ¹ | 185 | sup1 | Superscript 1 |
| º | 186 | ordm | Masculine ordinal |
| » | 187 | raquo | Right angle quotes |
| ¼ | 188 | frac14 | One quarter |
| ½ | 189 | frac12 | One half |
| ¾ | 190 | frac34 | Three quarters |
| ¿ | 191 | iquest | Inverted question mark |

| Char | Code | Name | Description |
|---|---|---|---|
| À | 192 | Agrave | A grave |
| Á | 193 | Aacute | A acute |
| Â | 194 | Acirc | A circumflex |
| Ã | 195 | Atilde | A tilde |
| Ä | 196 | Auml | A umlaut |
| Å | 197 | Aring | A ring |
| Æ | 198 | AElig | AE ligature |
| Ç | 199 | Ccedil | C cedilla |
| È | 200 | Egrave | E grave |
| É | 201 | Eacute | E acute |
| Ê | 202 | Ecirc | E circumflex |
| Ë | 203 | Euml | E umlaut |
| Ì | 204 | Igrave | I grave |
| Í | 205 | Iacute | I acute |
| Î | 206 | Icirc | I circumflex |
| Ï | 207 | Iuml | I umlaut |

| Char | Code | Name | Description |
|---|---|---|---|
| Ð | 208 | ETH | ETH |
| Ñ | 209 | Ntilde | N tilde |
| Ò | 210 | Ograve | O grave |
| Ó | 211 | Oacute | O acute |
| Ô | 212 | Ocirc | O circumflex |
| Õ | 213 | Otilde | O tilde |
| Ö | 214 | Ouml | O umlaut |
| × | 215 | times | Multiplication sign |
| Ø | 216 | Oslash | O slash |
| Ù | 217 | Ugrave | U grave |
| Ú | 218 | Uacute | U acute |
| Û | 219 | Ucirc | U circumflex |
| Ü | 220 | Uuml | U umlaut |
| Ý | 221 | Yacute | Y acute |
| Þ | 222 | THORN | THORN |
| ß | 223 | szlig | sharp s |

| Char | Code | Name | Description |
|---|---|---|---|
| à | 224 | agrave | a grave |
| á | 225 | aacute | a acute |
| â | 226 | acirc | a circumflex |
| ã | 227 | atilde | a tilde |
| ä | 228 | auml | a umlaut |
| å | 229 | aring | a ring |
| æ | 230 | aelig | ae ligature |
| ç | 231 | ccedil | c cedilla |
| è | 232 | egrave | e grave |
| é | 233 | eacute | e acute |
| ê | 234 | ecirc | e circumflex |
| ë | 235 | euml | e umlaut |
| ì | 236 | igrave | i grave |
| í | 237 | iacute | i acute |
| î | 238 | icirc | i circumflex |
| ï | 239 | iuml | i umlaut |

| Char | Code | Name | Description |
|---|---|---|---|
| ð | 240 | eth | eth |
| ñ | 241 | ntilde | n tilde |
| ò | 242 | ograve | o grave |
| ó | 243 | oacute | o acute |
| ô | 244 | ocirc | o circumflex |
| õ | 245 | otilde | o tilde |
| ö | 246 | ouml | o umlaut |
| ÷ | 247 | divide | division sign |
| ø | 248 | oslash | o slash |
| ù | 249 | ugrave | u grave |
| ú | 250 | uacute | u acute |
| û | 251 | ucirc | u circumflex |
| ü | 252 | uuml | u umlaut |
| ý | 253 | yacute | y acute |
| þ | 254 | thorn | thorn |
| ÿ | 255 | yuml | y umlaut |

# UTF-8

- Hoy en día usamos UTF-8.
- UTF-8 es idéntico a ASCII los primeros 128 Chars.
- UTF-8 es idéntico a ANSI y a 8859-1 de los valores 160 al 255.
- UTF-8 se sigue del valor 256 hasta el 10 000. Estos chars incluyen casi todos los lenguajes que se conocen. (Kanjis welcome!)

# HTML Encodings

- Para visualizar una página de HTML correctamente, el navegador debe conocer que codificación de caracteres se esta usando.

- Eso se define en el head a través del tag meta.

**<meta charset="UTF-8“>**