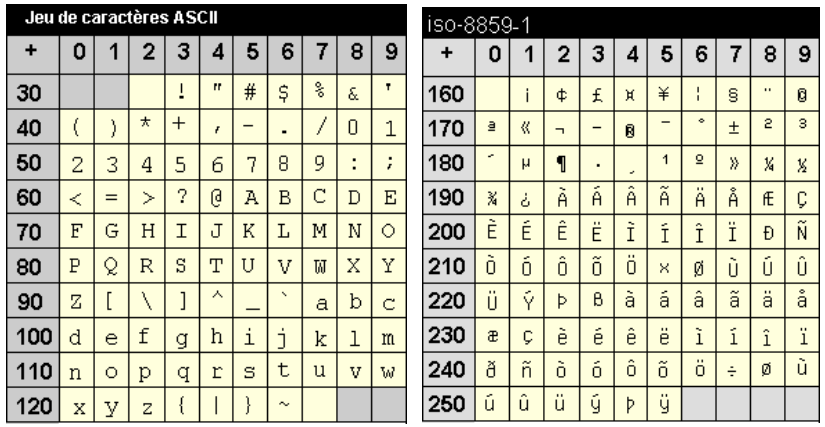
# HTML Encodage des caractères

- jeu de caractères (charset) ≠ encodage
- jeu de caractères:
  - liste de caractères identifiés avec un numéro unique (point de code)
  - exemples : ASCII , ISO-8859-1 (Latin-1), ISO-8859-5 (Cyrillique)…

ISO 8859-1 (Latin1)



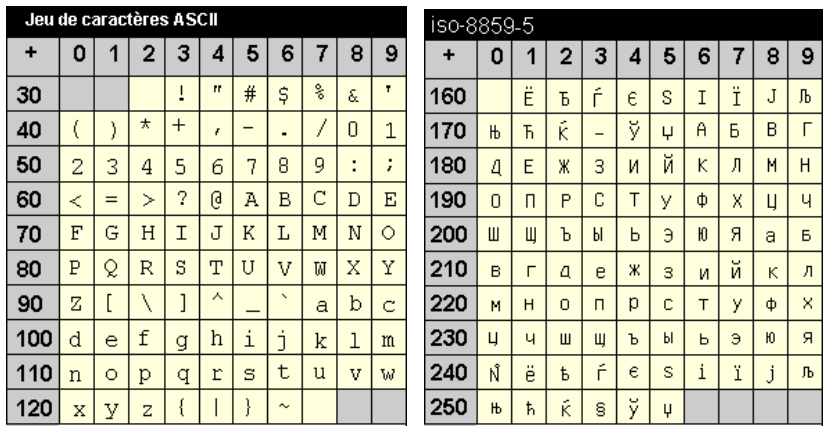
Jeu de caractères ASCII

| + | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 30 | | | | ! | " | # | $ | % | & | ' |
| 40 | ( | ) | * | + | , | - | . | / | 0 | 1 |
| 50 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | : | ; |
| 60 | < | = | > | ? | @ | A | B | C | D | E |
| 70 | F | G | H | I | J | K | L | M | N | O |
| 80 | P | Q | R | S | T | U | V | W | X | Y |
| 90 | Z | [ | \ | ] | ^ | _ | ` | a | b | c |
| 100 | d | e | f | g | h | i | j | k | l | m |
| 110 | n | o | p | q | r | s | t | u | v | w |
| 120 | x | y | z | { | \| | } | ~ | | | |

iso-8859-1

| + | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 160 | | ¡ | ¢ | £ | ¤ | ¥ | ¦ | § | ¨ | © |
| 170 | ª | « | ¬ | - | ® | ¯ | ° | ± | ² | ³ |
| 180 | ´ | µ | ¶ | · | ¸ | ¹ | º | » | ¼ | ½ |
| 190 | ¾ | ¿ | À | Á | Â | Ã | Ä | Å | Æ | Ç |
| 200 | È | É | Ê | Ë | Ì | Í | Î | Ï | Ð | Ñ |
| 210 | Ò | Ó | Ô | Õ | Ö | × | Ø | Ù | Ú | Û |
| 220 | Ü | Ý | Þ | ß | à | á | â | ã | ä | å |
| 230 | æ | ç | è | é | ê | ë | ì | í | î | ï |
| 240 | ð | ñ | ò | ó | ô | õ | ö | ÷ | ø | ù |
| 250 | ú | û | ü | ý | þ | ÿ | | | | |

ISO 8859-5 (Cyrillique)



Jeu de caractères ASCII

| + | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 30 | | | | ! | " | # | $ | % | & | ' |
| 40 | ( | ) | * | + | , | - | . | / | 0 | 1 |
| 50 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | : | ; |
| 60 | < | = | > | ? | @ | A | B | C | D | E |
| 70 | F | G | H | I | J | K | L | M | N | O |
| 80 | P | Q | R | S | T | U | V | W | X | Y |
| 90 | Z | [ | \ | ] | ^ | _ | ` | a | b | c |
| 100 | d | e | f | g | h | i | j | k | l | m |
| 110 | n | o | p | q | r | s | t | u | v | w |
| 120 | x | y | z | { | \| | } | ~ | | | |

iso-8859-5

| + | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 160 | | Ё | Ђ | Ѓ | Є | Ѕ | І | Ї | Ј | Љ |
| 170 | Њ | Ћ | Ќ | - | Ў | Џ | А | Б | В | Г |
| 180 | Д | Е | Ж | З | И | Й | К | Л | М | Н |
| 190 | О | П | Р | С | Т | У | Ф | Х | Ц | Ч |
| 200 | Ш | Щ | Ъ | Ы | Ь | Э | Ю | Я | а | б |
| 210 | в | г | д | е | ж | з | и | й | к | л |
| 220 | м | н | о | п | р | с | т | у | ф | х |
| 230 | ц | ч | ш | щ | ъ | ы | ь | э | ю | я |
| 240 | № | ё | ђ | ѓ | є | ѕ | і | ї | ј | љ |
| 250 | њ | ћ | ќ | § | ў | џ | | | | |

http://braesch.fr/sites/default/html/selfhtml/internationalisation/jeux_caracteres.htm

ANSI (Windows-1252) jeu de caractères de Windows. Identique à ISO-8859-1 (mis à part qu'ANSI a 32 caractères supplémentaires).

# HTML Encodage des caractères

- jeu de caractères (charset) ≠ encodage
- encodage de caractères
  - algorithme qui traduit un point de code en binaire

ISO 8859-1 (Latin1)

Jeu de caractères ASCII

| + | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 30 | | | | ! | " | # | $ | % | & | ' |
| 40 | ( | ) | * | + | , | - | . | / | 0 | 1 |
| 50 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | : | ; |
| 60 | < | = | > | ? | @ | A | B | C | D | E |
| 70 | F | G | H | I | J | K | L | M | N | O |
| 80 | P | Q | R | S | T | U | V | W | X | Y |
| 90 | Z | [ | \ | ] | ^ | _ | ` | a | b | c |
| 100 | d | e | f | g | h | i | j | k | l | m |
| 110 | n | o | p | q | r | s | t | u | v | w |
| 120 | x | y | z | { | \| | } | ~ | | | |

iso-8859-1

| + | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 160 | | ¡ | ¢ | £ | ¤ | ¥ | ¦ | § | ¨ | © |
| 170 | ª | « | ¬ | - | ® | ¯ | ° | ± | ² | ³ |
| 180 | ´ | µ | ¶ | · | ¸ | ¹ | º | » | ¼ | ½ |
| 190 | ¾ | ¿ | À | Á | Â | Ã | Ä | Å | Æ | Ç |
| 200 | È | É | Ê | Ë | Ì | Í | Î | Ï | Ð | Ñ |
| 210 | Ò | Ó | Ô | Õ | Ö | × | Ø | Ù | Ú | Û |
| 220 | Ü | Ý | Þ | ß | à | á | â | ã | ä | å |
| 230 | æ | ç | è | é | ê | ë | ì | í | î | ï |
| 240 | ð | ñ | ò | ó | ô | õ | ö | ÷ | ø | ù |
| 250 | ú | û | ü | ý | þ | ÿ | | | | |

ISO 8859-5

Jeu de caractères ASCII

| + | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 30 | | | | ! | " | # | $ | % | & | ' |
| 40 | ( | ) | * | + | , | - | . | / | 0 | 1 |
| 50 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | : | ; |
| 60 | < | = | > | ? | @ | A | B | C | D | E |
| 70 | F | G | H | I | J | K | L | M | N | O |
| 80 | P | Q | R | S | T | U | V | W | X | Y |
| 90 | Z | [ | \ | ] | ^ | _ | ` | a | b | c |
| 100 | d | e | f | g | h | i | j | k | l | m |
| 110 | n | o | p | q | r | s | t | u | v | w |
| 120 | x | y | z | { | \| | } | ~ | | | |

iso-8859-5

| + | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 160 | | Ё | Ђ | Ѓ | Є | Ѕ | І | Ї | Ј | Љ |
| 170 | Њ | Ћ | Ќ | - | Ў | Џ | А | Б | В | Г |
| 180 | Д | Е | Ж | З | И | Й | К | Л | М | Н |
| 190 | О | П | Р | С | Т | У | Ф | Х | Ц | Ч |
| 200 | Ш | Щ | Ъ | Ы | Ь | Э | Ю | Я | а | б |
| 210 | в | г | д | е | ж | з | и | й | к | л |
| 220 | м | н | о | п | р | с | т | у | ф | х |
| 230 | ц | ч | ш | щ | ъ | ы | ь | э | ю | я |
| 240 | № | ё | ђ | ѓ | є | ѕ | і | ї | ј | љ |
| 250 | њ | ћ | ќ | § | ў | џ | | | | |

A : 65 → $2^6 + 1$ → 0100 0001 → 4 1 (*binaire* → *héxadécimal*)

1110 1001 → $2^7 + 2^6 + 2^5 + 2^3 + 2^0$ → 233 (*binaire* → *héxadécimal*)

→ é ISO 8859-1

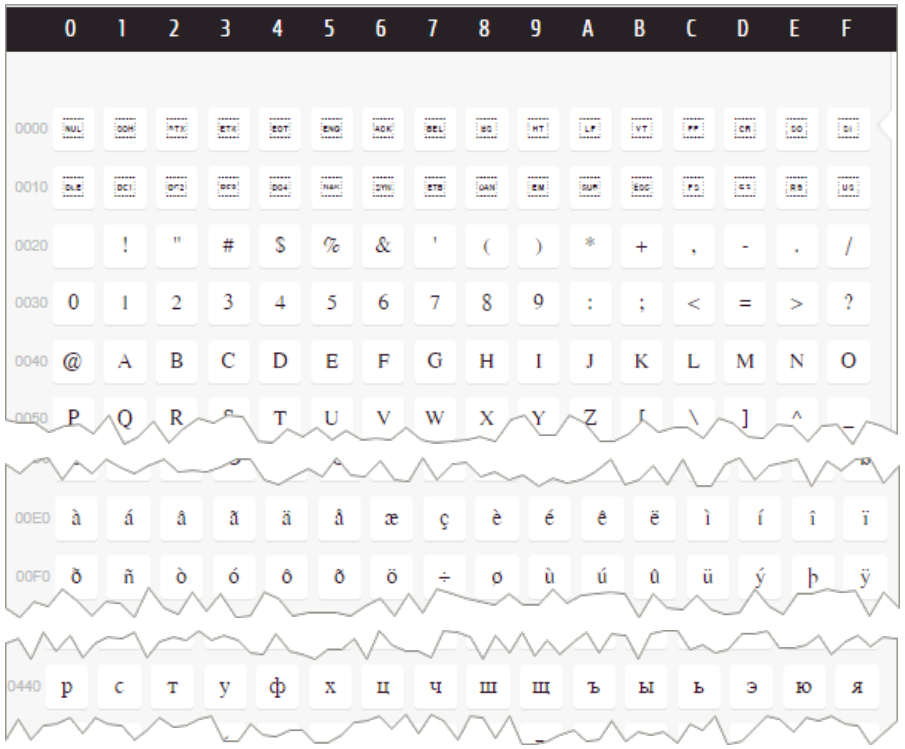
→ щ ISO 8859-5

# HTML Encodage des caractères

- caractères Unicodes jeu de caractère universel

**Unicode** est un standard informatique qui permet des échanges de textes dans différentes langues, à un niveau mondial. Il est développé par le *Consortium Unicode*, qui vise au codage de texte écrit en donnant à tout caractère de n'importe quel système d'écriture un nom et un identifiant numérique, et ce de manière unifiée, quelle que soit la plate-forme informatique ou le logiciel utilisés.

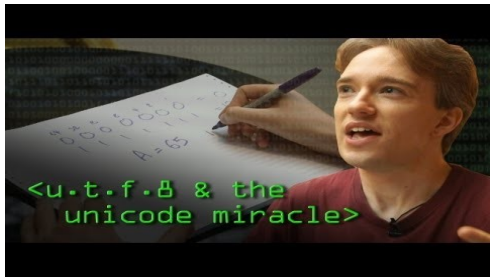
WIKIPEDIA
La enciclopedia libre




https://unicode-table.com/

**point de code**

| | décimal | hexadécimal |
|---|---|---|
| A | 65 | 0041 |
| é | 233 | 00E9 |
| щ | 1097 | 0449 |




https://www.youtube.com/watch?v=MijmeoH9LT4

# HTML Encodage des caractères

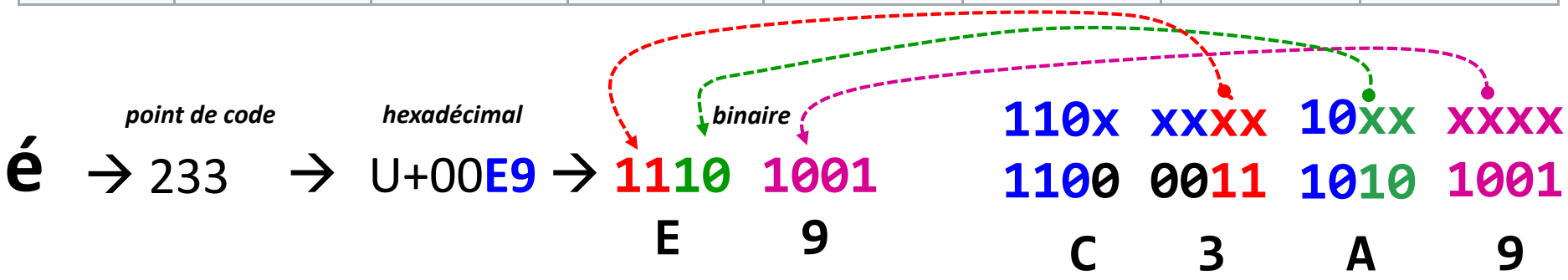
- caractères Unicode : un octet ne suffit plus
- encodages UTF *(Unicode Transformation Format)*
  - codage de taille variable
    - UTF-8 : 1 , 2 , 3  ou 4  octets
    - UTF-16 : 2 ou 4 octets  (codage des chaînes en Java)
  - codage de taille fixe
    - UTF-32: 4 octets

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

# HTML Encodage des caractères

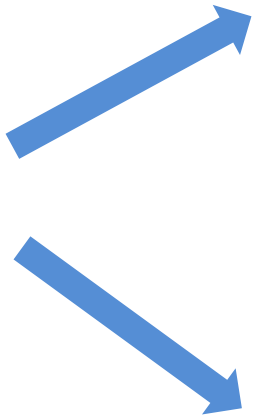
- encodages UTF-8
  - codage de taille variable
    - UTF-8 : 1 , 2 , 3 ou 4 octets

| Number of bytes | Bits for code point | First code point | Last code point | Byte 1 | Byte 2 | Byte 3 | Byte 4 |
|---|---|---|---|---|---|---|---|
| 1 | 7 | 0 U+0000 | 127 U+007F | 0xxxxxxx | | | |
| 2 | 11 | 128 U+0080 | 2047 U+07FF | 110xxxxx | 10xxxxxx | | |
| 3 | 16 | 2048 U+0800 | 65535 U+FFFF | 1110xxxx | 10xxxxxx | 10xxxxxx | |
| 4 | 21 | 65536 U+10000 | 1 114 111 U+10FFFF | 11110xxx | 10xxxxxx | 10xxxxxx | 10xxxxxx |



# HTML Encodage des caractères

```
<!DOCTYPE html>
<html>
    <head>
        <title>TEST</title>
    </head>
    <body>
        Vivement l'été
    </body>
</html>
```



UTF-8

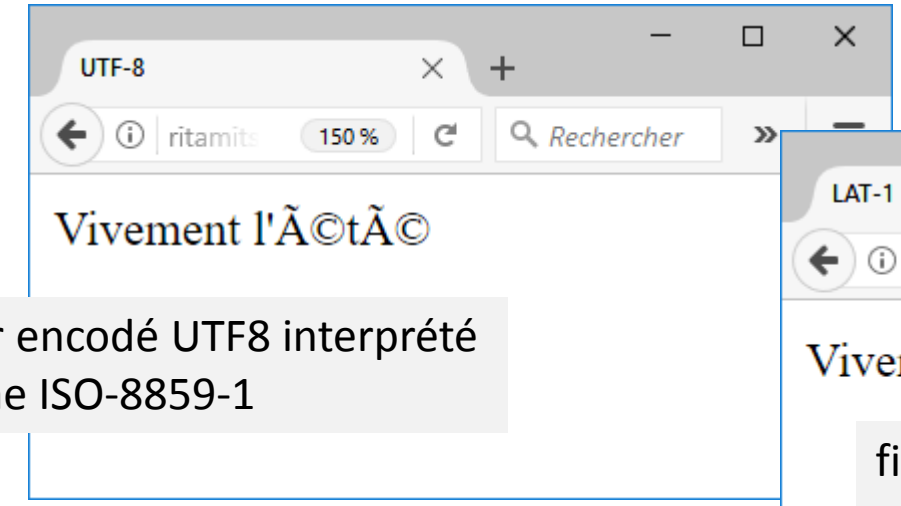
Lat-1
ISO 8859-1

```
$ od -t x1 testUTF8.html
0000000 3c 21 44 4f 43 54 59 50 45 20 68 74 6d 6c 3e 0d
0000020 0a 3c 68 74 6d 6c 3e 0d 0a 20 20 20 20 3c 68 65
0000040 61 64 3e 0d 0a 20 20 20 20 20 20 20 20 3c 74 69
0000060 74 6c 65 3e 54 45 53 54 3c 2f 74 69 74 6c 65 3e
0000100 0d 0a 20 20 20 20 3c 2f 68 65 61 64 3e 0d 0a 20
0000120 20 20 3c 62 6f 64 79 3e 0d 0a 20 20 20 20 20 20
0000140 20 20 20 56 69 76 65 6d 65 6e 74 20 6c 27 c3 a9
0000160 74 c3 a9 0d 0a 20 20 20 20 3c 2f 62 6f 64 79 3e
0000200 0d 0a 3c 2f 68 74 6d 6c 3e
0000211
```

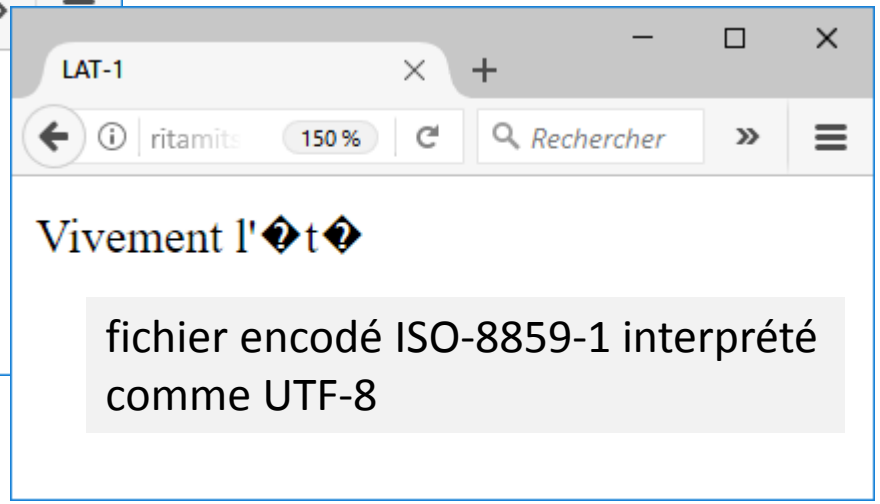
```
$ od -t x1 testLAT1.html
0000000 3c 21 44 4f 43 54 59 50 45 20 68 74 6d 6c 3e 0d
0000020 0a 3c 68 74 6d 6c 3e 0d 0a 20 20 20 20 3c 68 65
0000040 61 64 3e 0d 0a 20 20 20 20 20 20 20 20 3c 74 69
0000060 74 6c 65 3e 54 45 53 54 3c 2f 74 69 74 6c 65 3e
0000100 0d 0a 20 20 20 20 3c 2f 68 65 61 64 3e 0d 0a 20
0000120 20 20 3c 62 6f 64 79 3e 0d 0a 20 20 20 20 20 20
0000140 20 20 20 56 69 76 65 6d 65 6e 74 20 6c 27 e9 74
0000160 e9 0d 0a 20 20 20 20 3c 2f 62 6f 64 79 3e 0d 0a
0000200 3c 2f 68 74 6d 6c 3e
```

# HTML Encodage des caractères

- problèmes d'encodage
  - pas de déclaration d'encodage et l'encodage choisi par le navigateur n'est pas le bon
  - la déclaration d'encodage n'est pas la bonne




fichier encodé UTF8 interprété comme ISO-8859-1




fichier encodé ISO-8859-1 interprété comme UTF-8

pour comprendre menu encodage dans firefox
https://hsivonen.fi/encoding-menu

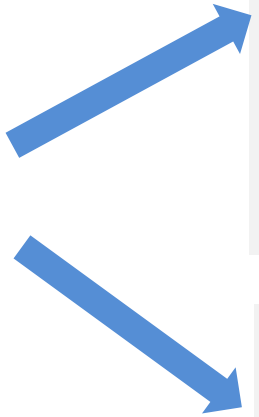
https://www.w3schools.com/html/html_charset.asp

https://lig-membres.imag.fr/genoud/teaching/PL2AI/cours/exemples/01_HTML/encodage/

Le serveur lig-membres.imag.fr ajoute un header `content-type` avec `charset='utf-8'` donc le navigateur interprète toutes les pages comme encodées en UTF-8 quelque soit la balise `<meta charset="…">`

# HTML Encodage des caractères

```
<!DOCTYPE html>
<html>
    <head>
        <title>TEST</title>
    </head>
    <body>
        Vivement l'été
    </body>
</html>
```



UTF-8

```
ritamitsouko:genoud% od -t x1 test-UTF8.html
0000000 3c 21 44 4f 43 54 59 50 45 20 68 74 6d 6c 3e 0d
0000020 0a 3c 68 74 6d 6c 3e 0d 0a 20 20 20 20 3c 68 65
0000040 61 64 3e 0d 0a 20 20 20 20 20 20 20 20 3c 74 69
0000060 74 6c 65 3e 54 45 53 54 3c 2f 74 69 74 6c 65 3e
0000100 0d 0a 20 20 20 20 3c 2f 68 65 61 64 3e 0d 0a 20
0000120 20 20 20 3c 62 6f 64 79 3e 0d 0a 20 20 20 20 20
0000140 20 20 20 56 69 76 65 6d 65 6e 74 20 6c 27 c3 a9
0000160 74 c3 a9 0d 0a 20 20 20 20 3c 2f 62 6f 64 79 3e
0000200 0d 0a 3c 2f 68 74 6d 6c 3e 0d 0a
0000213
```



Lat-1
ISO 8859-1

```
ritamitsouko:genoud%  od -t x1 testANSI.html
0000000 3c 21 44 4f 43 54 59 50 45 20 68 74 6d 6c 3e 0d
0000020 0a 3c 68 74 6d 6c 3e 0d 0a 20 20 20 20 3c 68 65
0000040 61 64 3e 0d 0a 20 20 20 20 20 20 20 20 3c 74 69
0000060 74 6c 65 3e 54 45 53 54 3c 2f 74 69 74 6c 65 3e
0000100 0d 0a 20 20 20 20 3c 2f 68 65 61 64 3e 0d 0a 20
0000120 20 20 20 3c 62 6f 64 79 3e 0d 0a 20 20 20 20 20
0000140 20 20 20 56 69 76 65 6d 65 6e 74 20 6c 27 e9 74
0000160 e9 0d 0a 20 20 20 20 3c 2f 62 6f 64 79 3e 0d 0a
0000200 3c 2f 68 74 6d 6c 3e 0d 0a
0000211
```