

Philippe GENOUD (LIG-STeamer)  
Philippe.Genoud@imag.fr

**M2 CCI - M2 GEOMAS - 2024-2025**  
**cours Technologies du Web (TW)**

# HTML et encodage des caractères

dernière modification : 30/09/2024 17:30



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).

# HTML Encodage des caractères

- du texte HTML peut être distribué sur le réseau en utilisant différents encodage des caractères.
- balise `<meta>` dans l'en tête indique au navigateur quel jeu de caractères (et encodage) est utilisé dans le document



```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN" "http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd">
<html>
  <head>
    <meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1">
    <title>LAT-1</title>
  </head>
  <body>
    Vivement l'été
  </body>
</html>
```



```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="UTF-8">
    <title>UTF-8</title>
  </head>
  <body>
    Vivement l'été
  </body>
</html>
```



Utiliser UTF-8 valeur recommandée par les standards du web

placer la spécification du **charset** au début de la partie **head**.

# HTML Encodage des caractères

- jeu de caractères (charset)  $\neq$  encodage
- jeu de caractères:
  - liste de caractères identifiés avec un numéro unique (point de code)
- encodage de caractères
  - algorithme qui traduit un point de code en binaire

Bits		Column													
b <sub>7</sub>	b <sub>6</sub>	b <sub>5</sub>	b <sub>4</sub>	b <sub>3</sub>	b <sub>2</sub>	b <sub>1</sub>	b <sub>0</sub>	0	1	2	3	4	5	6	7
0	0	0	0	0	0	0	0	NUL	DLE	SP	0	@	P	\	p
0	0	0	1	1	1	1	1	SOH	DC1	!	1	A	Q	a	q
0	0	1	0	0	0	0	0	STX	DC2	"	2	B	R	b	r
0	0	1	1	0	0	0	0	ETX	DC3	#	3	C	S	c	s
0	1	0	0	0	0	0	0	EOT	DC4	\$	4	D	T	d	t
0	1	0	1	0	0	0	0	ENQ	NAK	%	5	E	U	e	u
0	1	1	0	0	0	0	0	ACK	SYN	&	6	F	V	f	v
0	1	1	1	0	0	0	0	BEL	ETB	'	7	G	W	g	w
1	0	0	0	0	0	0	0	BS	CAN	(	8	H	X	h	x
1	0	0	1	0	0	0	0	HT	EM	)	9	I	Y	i	y
1	0	1	0	0	0	0	0	LF	SUB	*	:	J	Z	j	z
1	0	1	1	0	0	0	0	VT	ESC	+	;	K	[	k	{
1	1	0	0	0	0	0	0	FF	FS	,	<	L	\	l	
1	1	0	1	0	0	0	0	CR	GS	-	=	M	]	m	}
1	1	1	0	0	0	0	0	SO	RS	.	>	N	^	n	~
1	1	1	1	0	0	0	0	SI	US	/	?	O	_	o	DEL

ASCII chart from a pre-1972 printer manual  
<https://en.wikipedia.org/wiki/ASCII>

exemple : ASCII  
 (American Standard code for Information Interchange)

point code

66<sup>ème</sup> caractère du code ASCII

→ 65

⚠ on compte à partir de 0

A

encodage

4 1  
 $2^2 2^1 2^0$   $2^3 2^2 2^1 2^0$   
 100 0001

Hexadécimal  
 16 nombres représentés par 0 ... 9 A B C E F

Binaire

6 B  
 k 110 1101

$$65 = 2^6 + 2^0$$

# HTML Encodage des caractères

- Code ASCII insuffisant pour représenter les caractères de langues autres que l'anglais
- Définition d'autres jeux de caractères pour supporter différentes langues

ISO 8859-1 (Latin1)

Jeu de caractères ASCII										iso-8859-1											
+	0	1	2	3	4	5	6	7	8	9	+	0	1	2	3	4	5	6	7	8	9
30			!	"	#	\$	%	&	'		160		ı	ı	ı	ı	ı	ı	ı	ı	ı
40	(	)	*	+	,	-	.	/	0	1	170	ª	«	¬	-	»	-	°	±	²	³
50	2	3	4	5	6	7	8	9	:	;	180	ˆ	µ	¶	·	,	1	º	»	¼	½
60	<	=	>	?	@	A	B	C	D	E	190	¸	˘	ˆ	ˆ	ˆ	ˆ	ˆ	ˆ	ˆ	ˆ
70	F	G	H	I	J	K	L	M	N	O	200	È	É	Ê	Ë	Ì	Í	Î	Ï	Ð	Ñ
80	P	Q	R	S	T	U	V	W	X	Y	210	Õ	Ö	Ë	Ë	Ë	Ë	Ë	Ë	Ë	Ë
90	Z	[	\	]	^	_	`	a	b	c	220	Ü	Ý	Þ	ß	à	á	â	ã	ä	å
100	d	e	f	g	h	i	j	k	l	m	230	æ	ç	è	é	ê	ë	ì	í	î	ï
110	n	o	p	q	r	s	t	u	v	w	240	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù
120	x	y	z	{		}	~				250	ú	û	ü	ý	þ	ÿ				

ISO 8859-5 (Cyrillique)

Jeu de caractères ASCII										iso-8859-5											
+	0	1	2	3	4	5	6	7	8	9	+	0	1	2	3	4	5	6	7	8	9
30			!	"	#	\$	%	&	'		160		Ё	Ђ	Ѓ	Е	Ѕ	І	Ї	Ј	Љ
40	(	)	*	+	,	-	.	/	0	1	170	Њ	Ћ	Ќ	-	Ў	Ц	А	Б	В	Г
50	2	3	4	5	6	7	8	9	:	;	180	Д	Е	Ж	З	И	Й	К	Л	М	Н
60	<	=	>	?	@	A	B	C	D	E	190	О	П	Р	С	Т	У	Ф	Х	Ц	Ч
70	F	G	H	I	J	K	L	M	N	O	200	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я	а	б
80	P	Q	R	S	T	U	V	W	X	Y	210	в	г	д	е	ж	з	и	й	к	л
90	Z	[	\	]	^	_	`	a	b	c	220	м	н	о	п	р	с	т	у	ф	х
100	d	e	f	g	h	i	j	k	l	m	230	ч	ц	ш	щ	ъ	ы	ь	э	ю	я
110	n	o	p	q	r	s	t	u	v	w	240	Ѕ	ё	ђ	ѓ	е	ѕ	і	ї	ј	љ
120	x	y	z	{		}	~				250	њ	ќ	ќ	ѕ	ў	ц				

⚠ Attention  
contrairement  
à la table du code  
ASCII slide  
précédent, ces tables  
donnent les points de  
code et non pas  
l'encodage

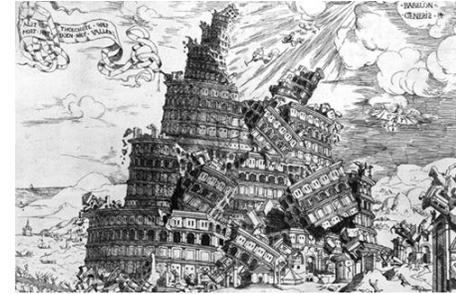
a → 97

[https://gilles-hunault.leria-info.univ-angers.fr/selfhtml\\_fr/internationalisation/jeux\\_caracteres.htm](https://gilles-hunault.leria-info.univ-angers.fr/selfhtml_fr/internationalisation/jeux_caracteres.htm)

ANSI (Windows-1252) jeu de caractères de Windows. Identique à ISO-8859-1 (certains caractères de contrôle plus utilisés ont été remplacés par des symboles utiles € , œ ...).

# HTML Encodage des caractères

- problèmes des jeux de caractères ISO
  - Prolifération des jeux de caractères : ISO-8859-2 (Latin-2) (langues d'Europe centrale : croate, serbe, ...), ISO-8859-3 (Latin-3) (galicien, le maltais, l'espéranto...) ....
  - Un même point de code et donc un même encodage peuvent correspondre à des caractères différents
  - Impossibilité de supporter les langages non alphabétiques (ex. idéogrammes chinois)



Cornelis Anthonisz (1547)

ISO 8859-1 (Latin1)

Jeu de caractères ASCII											iso-8859-1										
+	0	1	2	3	4	5	6	7	8	9	+	0	1	2	3	4	5	6	7	8	9
30			!	"	#	\$	%	&	'		160		ı	ı	ı	ı	ı	ı	ı	ı	ı
40	(	)	*	+	,	-	.	/	0	1	170	ª	«	¬	-	»	-	°	±	²	³
50	2	3	4	5	6	7	8	9	:	;	180	ˆ	µ	¶	·	,	1	º	»	¼	½
60	<	=	>	?	@	A	B	C	D	E	190	¸	˘	ˆ	ˆ	ˆ	ˆ	ˆ	ˆ	ˆ	ˆ
70	F	G	H	I	J	K	L	M	N	O	200	È	É	Ê	Ë	Ì	Í	Î	Ï	Ð	Ñ
80	P	Q	R	S	T	U	V	W	X	Y	210	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û
90	Z	[	\	]	^	_	`	a	b	c	220	Ü	Ý	Þ	ß	à	á	â	ã	ä	å
100	d	e	f	g	h	i	j	k	l	m	230	æ	ç	è	é	ê	ë	ì	í	î	ï
110	n	o	p	q	r	s	t	u	v	w	240	ä	ñ	ò	ó	ô	õ	ö	÷	ø	ù
120	x	y	z	{		}	~				250	ú	û	ü	ý	þ	ÿ				

ISO 8859-5 (Cyrilique)

Jeu de caractères ASCII											iso-8859-5										
+	0	1	2	3	4	5	6	7	8	9	+	0	1	2	3	4	5	6	7	8	9
30			!	"	#	\$	%	&	'		160		Ë	Ђ	Ѓ	Є	Ѕ	І	Ї	Ј	Љ
40	(	)	*	+	,	-	.	/	0	1	170	Њ	Ћ	Ќ	–	Ў	Џ	А	Б	В	Г
50	2	3	4	5	6	7	8	9	:	;	180	Д	Е	Ж	З	И	Й	К	Л	М	Н
60	<	=	>	?	@	A	B	C	D	E	190	О	П	Р	С	Т	У	Ф	Х	Ц	Ч
70	F	G	H	I	J	K	L	M	N	O	200	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я	а	б
80	P	Q	R	S	T	U	V	W	X	Y	210	в	г	д	е	ж	з	и	й	к	л
90	Z	[	\	]	^	_	`	a	b	c	220	м	н	о	п	р	с	т	у	ф	х
100	d	e	f	g	h	i	j	k	l	m	230	ч	ц	ш	щ	ъ	ы	ь	э	ю	я
110	n	o	p	q	r	s	t	u	v	w	240	Њ	ё	Ђ	Ѓ	Є	Ѕ	І	Ї	Ј	Љ
120	x	y	z	{		}	~				250	Њ	ћ	ќ	ѕ	џ	џ				

A : 65 → 2<sup>6</sup> + 1 → 0100 0001  
 4 1  
 hexadécimal

1110 1001 → 2<sup>7</sup> + 2<sup>6</sup> + 2<sup>5</sup> + 2<sup>3</sup> + 2<sup>0</sup> → 233  
 E 9  
 hexadécimal

→ é ISO 8859-1  
 → щ ISO 8859-5

# HTML Encodage des caractères

- caractères Unicode 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.



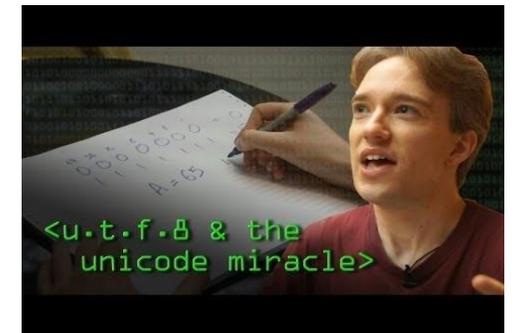
<https://home.unicode.org/about-unicode/>

v15 (sept 2022)  
149 186 caractères

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0000	NUL	SOH	STX	ETX	END	ACK	BS	HT	LF	VT	FF	CR	SO	SI		
0010	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	END	ESC	FS	GS	RS	US		
0020		!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
0030	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
0040	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
0050	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
00E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
00F0	ð	ñ	ò	ó	ô	õ	÷	ø	ù	ú	û	ü	ý	þ	ÿ	
0440	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я

<https://unicode-table.com/>

	point de code	
	décimal	hexadécimal
A	65	0041
é	233	00E9
Щ	1097	0449
😎	128 526	1F60E



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

# HTML Encodage des caractères

- caractères Unicode :
  - v 15.0 (sept. 2022) : 149 186 caractères ([https://unicode.org/faq/basic\\_q.html](https://unicode.org/faq/basic_q.html))
  - un octet ne suffit plus
- encodages UTF (*Unicode Transformation Format*)
  - codage de taille variable
    - UTF-8 : 1 , 2 , 3 ou 4 octets (préconisé pour le web)
    - UTF-16 : 2 ou 4 octets (codage des chaînes en Java et dans l'autres langages de programmation)
  - codage de taille fixe
    - UTF-32: 4 octets

[https://en.wikipedia.org/wiki/Comparison\\_of\\_Unicode\\_encodings](https://en.wikipedia.org/wiki/Comparison_of_Unicode_encodings)



# HTML Encodage des caractères

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

UTF-8

```
$ od -t x1 testUTF8.html
```

```
00000000 3c 21 44 4f 43 54 59 50 45 20 68 74 6d 6c 3e 0d
00000200 0a 3c 68 74 6d 6c 3e 0d 0a 20 20 20 20 3c 68 65
00000400 61 64 3e 0d 0a 20 20 20 20 20 20 20 20 3c 74 69
00000600 74 6c 65 3e 54 45 53 54 3c 2f 74 69 74 6c 65 3e
00001000 0d 0a 20 20 20 20 3c 2f 68 65 61 64 3e 0d 0a 20
00001200 20 20 20 3c 62 6f 64 79 3e 0d 0a 20 20 20 20 20
00001400 20 20 20 56 69 76 65 6d 65 6e 74 20 6c 27 c3 a9
00001600 74 c3 a9 0d 0a 20 20 20 20 3c 2f 62 6f 64 79 3e
00002000 0d 0a 3c 2f 68 74 6d 6c 3e
00002111
```

Lat-1  
ISO 8859-1

```
$ od -t x1 testLAT1.html
```

```
00000000 3c 21 44 4f 43 54 59 50 45 20 68 74 6d 6c 3e 0d
00000200 0a 3c 68 74 6d 6c 3e 0d 0a 20 20 20 20 3c 68 65
00000400 61 64 3e 0d 0a 20 20 20 20 20 20 20 20 3c 74 69
00000600 74 6c 65 3e 54 45 53 54 3c 2f 74 69 74 6c 65 3e
00001000 0d 0a 20 20 20 20 3c 2f 68 65 61 64 3e 0d 0a 20
00001200 20 20 20 3c 62 6f 64 79 3e 0d 0a 20 20 20 20 20
00001400 20 20 20 56 69 76 65 6d 65 6e 74 20 6c 27 e9 74
00001600 e9 0d 0a 20 20 20 20 3c 2f 62 6f 64 79 3e 0d 0a
00002000 3c 2f 68 74 6d 6c 3e
```

# HTML Encodage des caractères

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

UTF-8

```
$ od -t x1 testUTF8.html
```

```
00000000 3c 21 44 4f 43 54 59 50 45 20 68 74 6d 6c 3e 0d
00000020 0a 3c 68 74 6d 6c 3e 0d 0a 20 20 20 20 3c 68 65
00000040 61 64 3e 0d 0a 20 20 20 20 20 20 20 20 3c 74 69
00000060 74 6c 65 3e 54 45 53 54 3c 2f 74 69 74 6c 65 3e
00001000 0d 0a 20 20 20 20 3c 2f 68 65 61 64 3e 0d 0a 20
00001200 20 20 20 3c 62 6f 64 79 3e 0d 0a 20 20 20 20 20
00001400 20 20 20 56 69 76 65 6d 65 6e 74 20 6c 27 c3 a9
00001600 74 c3 a9 0d 0a 20 20 20 20 3c 2f 62 6f 64 79 3e
00002000 0d 0a 3c 2f 68 74 6d 6c 3e
00002111
```

Lat-1

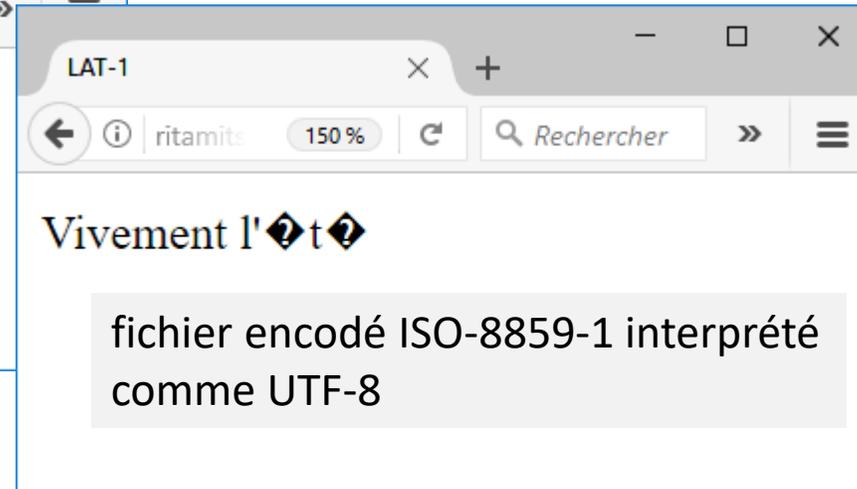
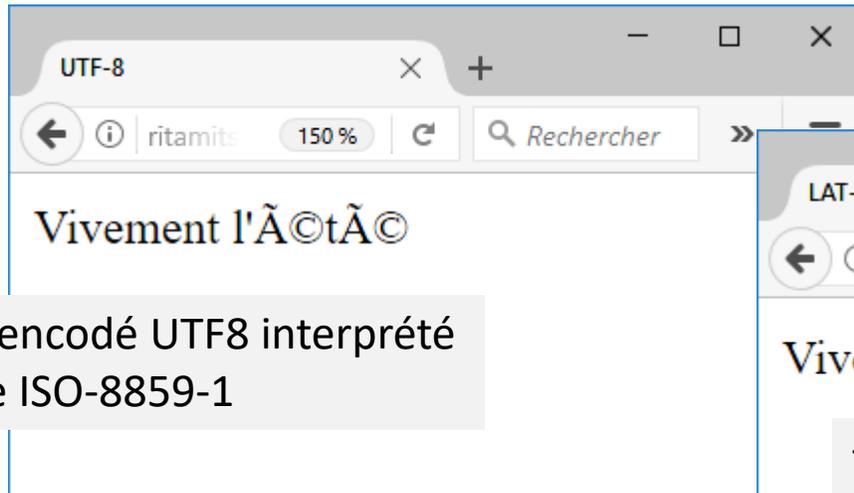
ISO 8859-1

```
$ od -t x1 testLAT1.html
```

```
00000000 3c 21 44 4f 43 54 59 50 45 20 68 74 6d 6c 3e 0d
00000020 0a 3c 68 74 6d 6c 3e 0d 0a 20 20 20 20 3c 68 65
00000040 61 64 3e 0d 0a 20 20 20 20 20 20 20 20 3c 74 69
00000060 74 6c 65 3e 54 45 53 54 3c 2f 74 69 74 6c 65 3e
00001000 0d 0a 20 20 20 20 3c 2f 68 65 61 64 3e 0d 0a 20
00001200 20 20 20 3c 62 6f 64 79 3e 0d 0a 20 20 20 20 20
00001400 20 20 20 56 69 76 65 6d 65 6e 74 20 6c 27 e9 74
00001600 e9 0d 0a 20 20 20 20 3c 2f 62 6f 64 79 3e 0d 0a
00002000 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



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

[https://www.w3schools.com/html/html\\_charset.asp](https://www.w3schools.com/html/html_charset.asp)

[https://lig-membres.imag.fr/genoud/teaching/PL2AI/cours/exemples/01\\_HTML/encodage/](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 quelle que 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
```