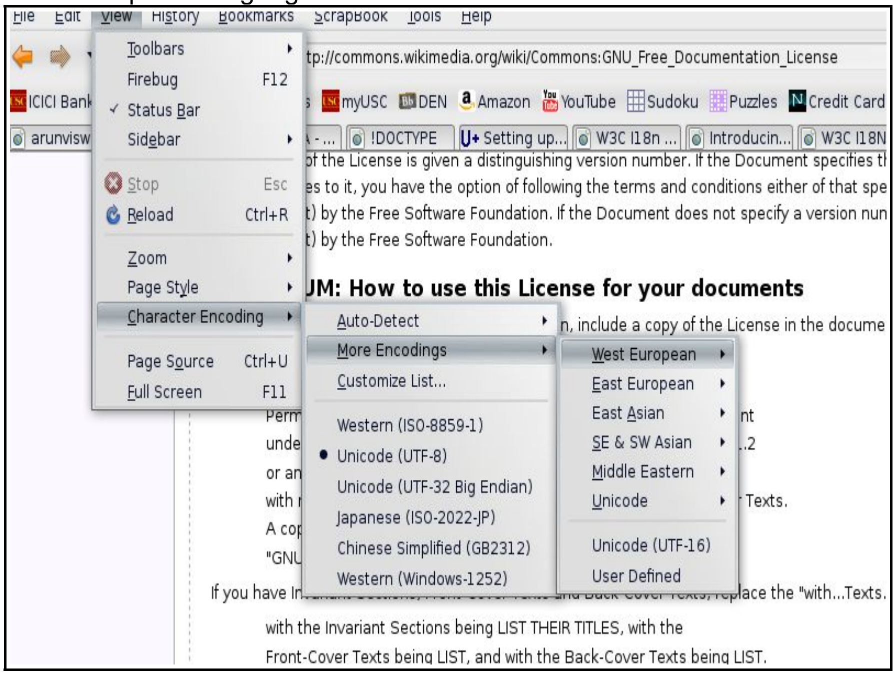
How browsers support Multilingual Content

Arun Viswanathan (aviswana@ieee.org) 01/23/09

Multilingual content is supported using internationalization (i18n) support in browsers. In simple terms its the support for displaying information in various languages. Every language is supported by specifying its characters as a **Coded Character Set**. The coded character set has a set of characters with a unique number (called a **code point**) assigned to each character. For example, English will have a character set, Hindi will have a character set and so on. The character 'A' in English has the unique identifier (or code point) of 65. **Character encoding** is the way these abstract characters are mapped to bytes for manipulation in a computer. For example, the character 'A' is represented in a byte in the ISO-8859-1 encoding while it can be represented in 1, 2, or 4 bytes in the Unicode Encoding Formats. It is important to note the distinction between the character set and its encoding. *There is only one character set but there can be multiple encodings.*

For XML and HTML (from version 4.0 onwards) the document character set is defined to be the Universal Character Set (UCS) or Unicode Standard. What this means is that HTML and XML documents will *use the Unicode Character Set but can have any Character Encodings.[1]*

There are many Character Encodings in use today. One easy way to find them out is to open the Firefox browser and look under View → Character Encodings. It should be noted that the Unicode Character Set and its associated encodings (UTF-8/UTF-16/UTF-32) are supposed to encompass all the existing character encodings in future. Also, what Unicode allows is to use multiple languages on the same page (say Hebrew, English and Chinese). This is something which is difficult to support with say the ISO-8559 series of encodings because they only encode the Western European Languages.


File Edit View History Bookmarks ScrapBook Tools Help
tp://commons.wikimedia.org/wiki/Commons:GNU_Free_Documentation_License
ICICI Bank
myUSC
DEN
Amazon
YouTube
Sudoku
Puzzles
Credit Card
arunvisw
!DOCTYPE
Setting up...
W3C I18n ...
Introducin...
W3C I18N
Toolbars
Firebug F12
✓ Status Bar
Sidebar
Stop Esc
Reload Ctrl+R
Zoom
Page Style
Character Encoding
Page Source Ctrl+U
Full Screen F11
Auto-Detect
More Encodings
Customize List...
Western (ISO-8859-1)
● Unicode (UTF-8)
Unicode (UTF-32 Big Endian)
Japanese (ISO-2022-JP)
Chinese Simplified (GB2312)
Western (Windows-1252)
West European
East European
East Asian
SE & SW Asian
Middle Eastern
Unicode
Unicode (UTF-16)
User Defined
of the License is given a distinguishing version number. If the Document specifies th
es to it, you have the option of following the terms and conditions either of that spe
t) by the Free Software Foundation. If the Document does not specify a version num
t) by the Free Software Foundation.
UM: How to use this License for your documents
n, include a copy of the License in the docume
If you have In... replace the "with...Texts.
with the Invariant Sections being LIST THEIR TITLES, with the
Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

These are ways in which one can define the encodings for HTML/XML/CSS pages.

1) Define encoding in the HTTP Header sent by the server

```
Content-Type: text/html; charset=iso-8859-1
```

2) For HTML or XHTML document (with content-type: text/html)

```
<meta http-equiv="Content-type" content="text/html; charset=UTF-8"/>
<link href="okinawa.html" rel="parent" charset="euc-jp"/>
```

3) For XML

```
<?xml version="1.0" encoding="UTF-8"?>
```

4) For CSS Style Sheets

```
@charset "utf-8";
```

Values for the encoding attribute can be found in the IANA registry [3]. Though these are called *charset*, they refer to the encodings and not the character sets.

These are the Precedence Rules that browsers will use if there are encodings specified in multiple places.

1. HTTP header Content-Type
2. XML declaration
3. <meta … charset=...> declaration
4. <link … charset=...> declaration

For external, linked CSS style sheets the precedence rules are:

1. HTTP header Content-Type
2. @charset rule
3. <link charset=".." rel="stylesheet" … />

Experiments using Firefox

Loading the website http://www.xinhuanet.com automatically displays the page in Chinese characters. This is because the encoding has been defined correctly by the authors of the page and my firefox browser understands the encoding.

HTTP capture of one of the GET requests shows that the response (in red) contains the charset GB2312. This helps the browser in identifying the encoding used in the page.

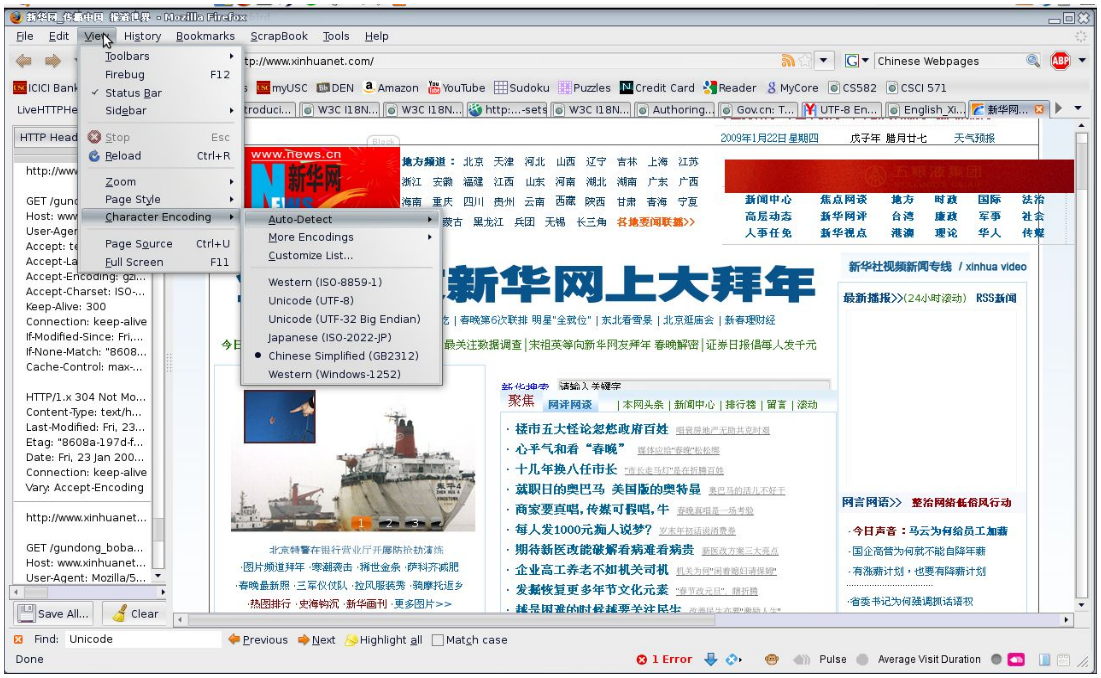
```
http://www.xinhuanet.com/gundong_bobao.htm
GET /gundong_bobao.htm HTTP/1.1
Host: www.xinhuanet.com
```

```
User-Agent: Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.9.0.5) Gecko/2008121622
Ubuntu/8.10 (intrepid) Firefox/3.0.5
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-us,en;q=0.5
Accept-Encoding: gzip,deflate
Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7
Keep-Alive: 300
Connection: keep-alive
If-Modified-Since: Fri, 23 Jan 2009 05:51:33 GMT
If-None-Match: "8608a-197d-fb30c340"
Cache-Control: max-age=0

HTTP/1.x 304 Not Modified
Content-Type: text/html; charset=GB2312
Last-Modified: Fri, 23 Jan 2009 05:51:33 GMT
Etag: "8608a-197d-fb30c340"
Date: Fri, 23 Jan 2009 05:49:34 GMT
Connection: keep-alive
Vary: Accept-Encoding
```

Additionally, the source code of the page reveals that the page contains the following meta tag.

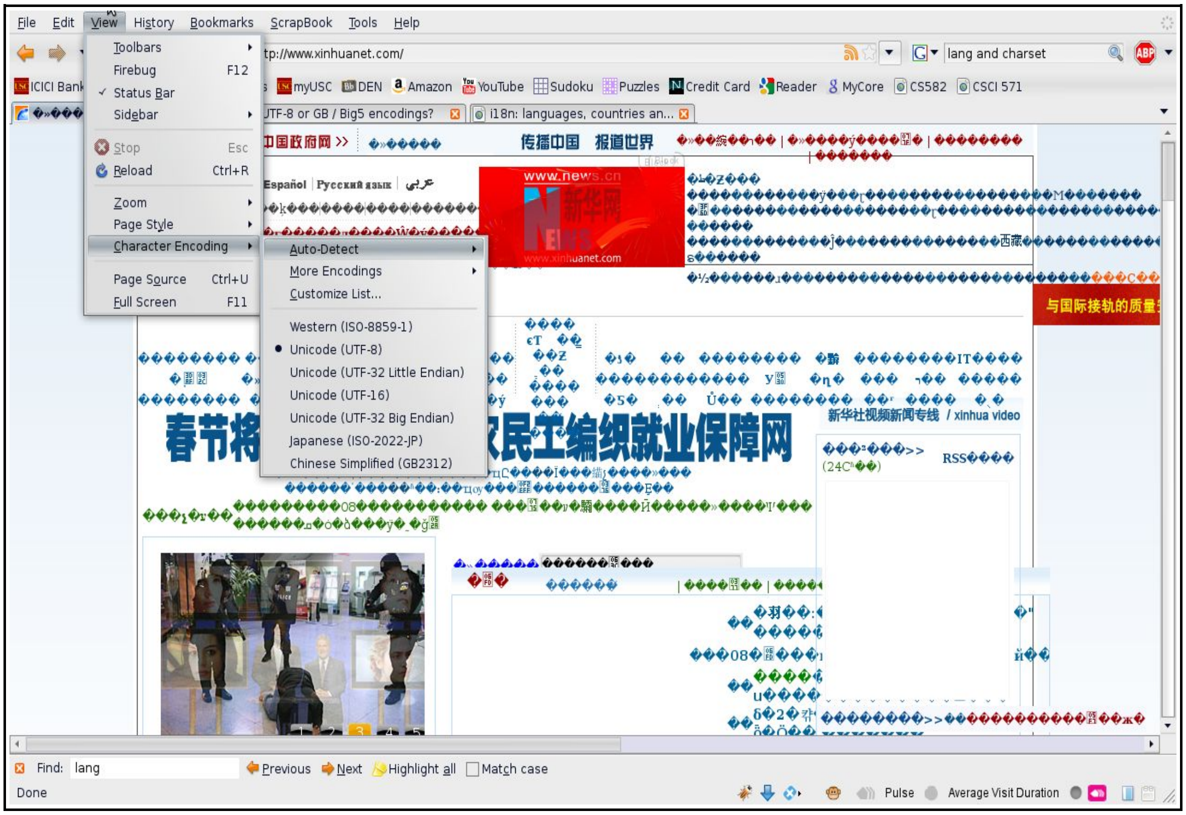
```
<meta http-equiv="Content-Type" content="text/html; charset=gb2312" />
```

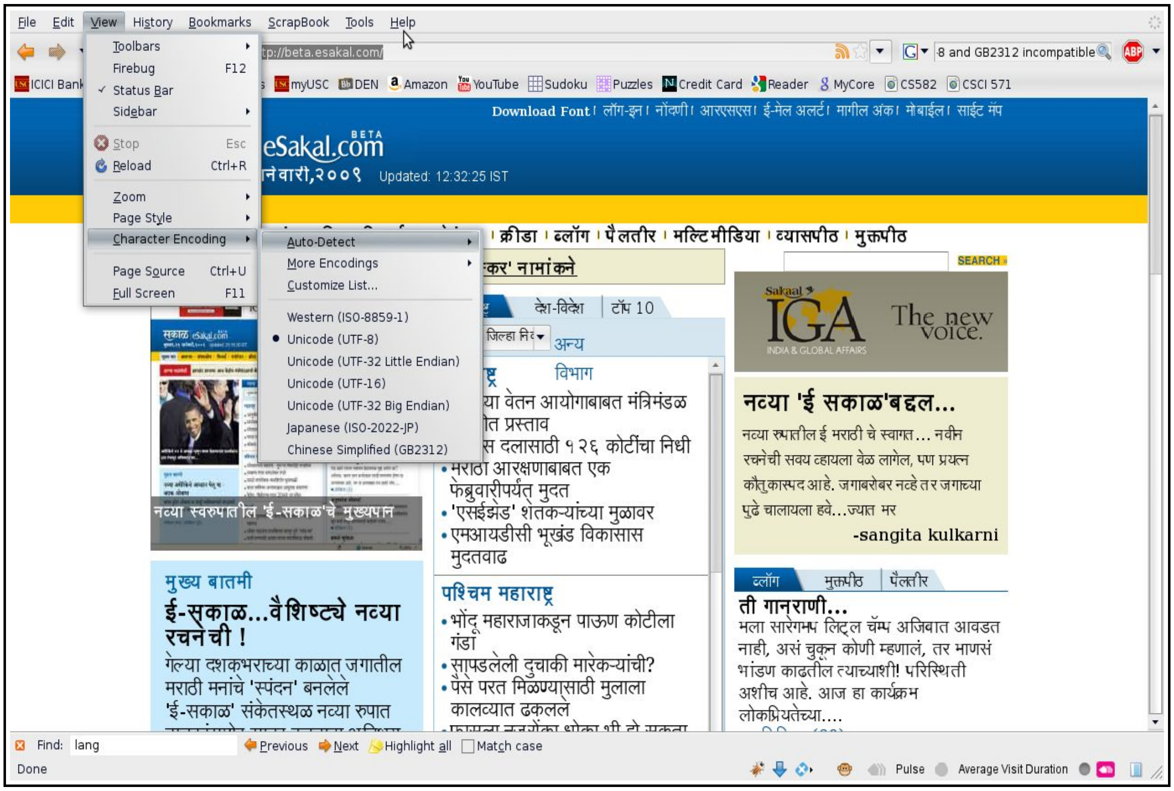
Also, firefox automatically displays the current character encoding that it is using under View → Character Encoding. As you can see below, the encoding is displayed as Chinese Simplified (GB2312)



```
http://www...

GET /gund...
Host: www...
User-Ager...
Accept: te...
Accept-La...
Accept-Encoding: gzi...
Accept-Charset: ISO-...
Keep-Alive: 300
Connection: keep-alive
If-Modified-Since: Fri,...
If-None-Match: "8608...
Cache-Control: max-...

HTTP/1.x 304 Not Mo...
Content-Type: text/h...
Last-Modified: Fri, 23...
Etag: "8608a-197d-f...
Date: Fri, 23 Jan 200...
Connection: keep-alive
Vary: Accept-Encoding

http://www.xinhuanet...

GET /gundong_boba...
Host: www.xinhuanet...
User-Agent: Mozilla/5...
```

Another experiment I did was to change the encoding to UTF-8 by using *View → Character Encodings* and selecting UTF-8. Unexpectedly, the page turned to gibberish. According to definitions, the document is supposed to correctly display in UTF-8 too but it does not. It turns out the UTF-8 (inspite of its Chinese support) may not be the most efficient encodings for Chinese fonts. The GB2312 encoding is more efficient (according to [4]) and hence is the preferred choice of encoding in mainland China.


File Edit View History Bookmarks ScrapBook Tools Help
tp://www.xinhuanet.com/
lang and charset
Toolbars
Firebug F12
✓ Status Bar
Sidebar
Stop Esc
Reload Ctrl+R
Zoom
Page Style
Character Encoding
Page Source Ctrl+U
Full Screen F11
Auto-Detect
More Encodings
Customize List...
Western (ISO-8859-1)
● Unicode (UTF-8)
Unicode (UTF-32 Little Endian)
Unicode (UTF-16)
Unicode (UTF-32 Big Endian)
Japanese (ISO-2022-JP)
Chinese Simplified (GB2312)
ICICI Bank myUSC DEN Amazon YouTube Sudoku Puzzles Credit Card Reader MyCore CS582 CSCI 571
UTF-8 or GB / Big5 encodings?
i18n: languages, countries an...
传播中国 报道世界
www.news.cn 新华网
春节将 农民工编织就业保障网
新华社视频新闻专线 / xinhua video
与国际接轨的质量
Find: lang Previous Next Highlight all Match case
Done
Pulse Average Visit Duration


File Edit View History Bookmarks ScrapBook Tools Help
Toolbars
Firebug F12
✓ Status Bar
Sidebar
Stop Esc
Reload Ctrl+R
Zoom
Page Style
Character Encoding
Page Source Ctrl+U
Full Screen F11
Auto-Detect
More Encodings
Customize List...
Western (ISO-8859-1)
● Unicode (UTF-8)
Unicode (UTF-32 Little Endian)
Unicode (UTF-16)
Unicode (UTF-32 Big Endian)
Japanese (ISO-2022-JP)
Chinese Simplified (GB2312)
tp://beta.esakal.com/
8 and GB2312 incompatible
ICICI Bank myUSC DEN Amazon YouTube Sudoku Puzzles Credit Card Reader MyCore CS582 CSCI 571
Download Font। लॉग-इन। नोंदणी। आरएसएस। ई-मेल अलर्ट। मागील अंक। मोबाईल। साईट मॅप
eSakal.com BETA
नेवारी,२००९ Updated: 12:32:25 IST
क्रीडा । ब्लॉग । पैलतीर । मल्टिमीडिया । व्यासपीठ । मुक्तपीठ
SEARCH »
Sakaal IGA INDIA & GLOBAL AFFAIRS The new voice.
देश-विदेश टॉप 10
जिल्हा निवडा अन्य
विभाग
नव्या स्वरुपातील 'ई-सकाळ'चे मुख्यपान
ब्लॉग मुक्तपीठ पैलतीर
Find: lang Previous Next Highlight all □ Match case
Done Pulse Average Visit Duration

मुख्य बातमी

ई-सकाळ...वैशिष्ट्ये नव्या रचनेची !

गेल्या दशकभराच्या काळात जगातील मराठी मनांचे 'स्पंदन' बनलेले 'ई-सकाळ' संकेतस्थळ नव्या रुपात

- मराठा आरक्षणाबाबत एक फेब्रुवारीपर्यंत मुदत
- 'एसईझेड' शेतकऱ्यांच्या मुळावर
- एमआयडीसी भूखंड विकासास मुदतवाढ

पश्चिम महाराष्ट्र

- भोंदू महाराजाकडून पाऊण कोटीला गंडा
- सापडलेली दुचाकी मारेकऱ्यांची?
- पैसे परत मिळण्यासाठी मुलाला कालव्यात ढकलले

नव्या 'ई सकाळ'बद्दल...

नव्या रूपातील ई मराठी चे स्वागत... नवीन रचनेची सवय व्हायला वेळ लागेल, पण प्रयत्न कौतुकास्पद आहे. जगाबरोबर नव्हे तर जगाच्या पुढे चालायला हवे...ज्यात मर

-sangita kulkarni

ती गानराणी...

मला सारेगमप लिट्ल चॅम्प अजिबात आवडत नाही, असं चुकून कोणी म्हणालं, तर माणसं भांडण काढतील त्याच्याशी! परिस्थिती अशीच आहे. आज हा कार्यक्रम लोकप्रियतेच्या....

An Indian website (shown above) in the Marathi language uses the UTF-8 encoding by default. This ensures maximal compatibility as every browser would at the least support the UTF-8 encoding.

Supporting Multilingual options on the same page

There can be two possible solutions:

1) Use different URLs for multiple languages and put them on top of the page. Like for example, xinhuanet gives you the same news in English/French/Russian/Arabic/Latin etc. Clicking on the links takes you to a different webpage where the encodings are specified differently.

Something interesting i found out about this after clicking on each link:

Language	Encoding in HTTP header	Encoding specified in Index.html
Chinese (GB) Version	GB2312	GB2312
Chinese (Big5) Version	Big5	Big5
English	**GB2312**	**ISO-8859-1**
Russian	utf-8	utf-8
Arabic	utf-8	utf-8
French	**Utf-8**	**iso-8859-1**
Latin	Utf-8	utf-8

There is a discrepancy in the encodings specified for the English and French sites. And according to priority the encoding specified in the HTTP header would be used. But it seems GB2312 also supports the basic ASCII set and hence the English version still displays correctly. But i think this is a bug and cause problems.

2) The other solution is to use UTF-8 as the default encoding as it allows one to use a combination of character sets (thats what it is for). One could use javascript to hide display information according to language selected. This is the recommended way i guess.

References

[1] http://www.w3.org/International/tutorials/tutorial-char-enc/

[2] http://www.w3.org/International/geo/html-tech/tech-character.html
[3] http://www.iana.org/assignments/character-sets
[4] http://en.wikipedia.org/wiki/GB2312