



WHITE PAPER



# MULTI-BYTE CHARACTER SETS

Reporting Data in Multiple Languages

August 27, 2008

Version 2.0





# 1 Multi-Byte Character Sets (Internationalization)

SiteCatalyst allows data to be captured and reported in multiple languages, which allows international sites to be easily tagged with SiteCatalyst code, and generate reports that reflect the site content as displayed to the user. A single report suite can be used to collect and report data in multiple languages.

Properly utilizing the internationalization capability of SiteCatalyst involves coordination of the report suite configuration, web page encoding and the SiteCatalyst property *charSet*. For example, if the sites *mysite.com* (English), *mysite.co.jp* (Japanese) and *mysite.co.kr* (Korean) are all sending data to a single global report suite, SiteCatalyst can display the English, Japanese, and Korean data simultaneously in a single report.

In addition to collecting and displaying international data, the SiteCatalyst interface can be displayed in several languages, including English, German, Japanese, Chinese, and Korean.

## 1.1 Web Page Encodings and Character Sets

Web pages display textual data by converting numeric character codes to physical characters based on the page encoding, which defines the range of available characters that can be properly displayed on the page. The page encoding is set with one of the following methods.

- Using a <META> tag inside the <HEAD> tag of the page, for example, <META http-equiv="Content-Type" content="text/html;charSet=ISO-8859-1">
- Within the http header, for example, Content-Type: text/html; charSet=ISO-8859-1
- By browser auto-detection; If methods 1 and 2 are not used, modern browsers will attempt to detect the page encoding based on the content or simply use a default encoding based on user preferences.

For greater visibility of the page encoding, Omniture recommends using the first whenever possible. The third method may be unreliable for international sites and should be avoided whenever possible.

For additional information on encodings and character sets, refer to <http://www.w3.org/International/tutorials/tutorial-char-enc/>.

### 1.1.1 ISO-8859-1 Encoding and Character Set

The most commonly used encoding for Latin based languages (English, French, Spanish, etc.) is "ISO-8859-1," which is one of many standards that use single-byte encodings. Each character is represented by one (and only one) byte of data. Therefore, single-byte encodings, including ISO-8859-1, is limited to 256 displayable characters. The table below shows the complete set of characters that are available within ISO-8859-1.

**Table 1-A: ISO 8859-1 Character Set**

0-31 non-displayed control codes			128-159 unused								
32	space	64	@	96	`	160	space	192	À	224	à
33	!	65	A	97	a	161	ı	193	Á	225	á
34	"	66	B	98	b	162	ç	194	Â	226	â
35	#	67	C	99	c	163	£	195	Ã	227	ã
36	`	68	D	100	d	164	¤	196	Ä	228	ä

37	%	69	E	101	e	165	¥	197	À	229	ã
38	&	70	F	102	f	166	¦	198	Æ	230	æ
39	'	71	G	103	g	167	§	199	Ç	231	ç
40	(	72	H	104	h	168	¨	200	È	232	è
41	)	73	I	105	i	169	©	201	É	233	é
42	*	74	J	106	j	170	ª	202	Ê	234	ê
43	+	75	K	107	k	171	«	203	Ë	235	ë
44	,	76	L	108	l	172	¬	204	Ì	236	ì
45	-	77	M	109	m	173	s hyphen	205	Í	237	í
46	.	78	N	110	n	174	®	206	Î	238	î
47	/	79	O	111	o	175	¯	207	Ï	239	ï
48	0	80	P	112	p	176	°	208	Ð	240	ð
49	1	81	Q	113	q	177	±	209	Ñ	241	ñ
50	2	82	R	114	r	178	²	210	Ò	242	ò
51	3	83	S	115	s	179	³	211	Ó	243	ó
52	4	84	T	116	t	180	´	212	Ô	244	ô
53	5	85	U	117	u	181	µ	213	Õ	245	õ
54	6	86	V	118	v	182	¶	214	Ö	246	ö
55	7	87	W	119	w	183	·	215	×	247	÷
56	8	88	X	120	x	184	,	216	Ø	248	ø
57	9	89	Y	121	y	185	¹	217	Ù	249	ù
58	:	90	Z	122	z	186	º	218	Ú	250	ú
59	;	91	[	123	{	187	»	219	Û	251	û
60	<	92	\	124		188	¼	220	Ü	252	ü
61	=	93	]	125	}	189	½	221	Ý	253	ý
62	>	94	^	126	~	190	¾	222	Þ	254	þ
63	?	95	_	127	delete	191	¿	223	ß	255	ÿ

### 1.1.2 The CP1252 (Windows-1252) Character Set

The CP1252 encoding and character set (otherwise known as the Windows-1252 or simply Windows character set) is a superset of ISO-8859-1. The CP1252 character set was developed by Microsoft and is used primarily by Microsoft Windows systems. This encoding uses the 128-159 code range to display additional characters not included in the ISO-8859-1 character set.

**Table 1-B: CP1252 Character Set**

128	€	134	†	140	Œ	146	'	152	~	158	ž
129		135	‡	141		147	“	153	™	159	ÿ
130	,	136	^	142	Ž	148	”	154	š		
131	f	137	[	143		149	•	155	›		
132	„	138	Š	144		150	–	156	œ		
133	...	139	<	145	‘	151	—	157			

Since this character set is not standardized across all platforms and browsers, these character codes are not valid HTML, though they will display properly on some systems and browsers. Use of these character codes will result in inconsistent display across browser versions and operating systems. To properly display these characters requires a more advanced character set and encoding, such as the Unicode based UTF-8 (see below).

### 1.1.3 UTF-8 Encoding (Unicode Character Set)

UTF-8 encoding is quickly becoming the standard for displaying multilingual (as well as mathematical and scientific) data on the web. UTF-8 is based on the standardized (but evolving) Unicode character set. Unicode is an advanced character set that as of version 4.0, includes over 70,000 characters from nearly all written languages. UTF-8 is one of the most common encoding methods used to convert Unicode character codes into a data byte sequence. Unlike single-byte encoding methods, each character can consist of one to four bytes of data in Unicode. For more information on Unicode and UTF-8, refer to the following web sites.

- <http://www.unicode.org>
- <http://en.wikipedia.org/wiki/Unicode>
- <http://en.wikipedia.org/wiki/UTF-8>

## 1.2 SiteCatalyst Report Suites (Standard ISO and Multi-byte Enabled)

Each SiteCatalyst report suite is configured to be either standard (or ISO) or a multi-byte (UTF-8/localized) report suite. This setting determines what encoding is to be used to store and display SiteCatalyst data. A standard report suite uses ISO-8859-1 encoding while a multi-byte suite uses UTF-8 encoding. Any characters that are not in the ISO-8859-1 character set (including those in the CP 1252 character set) will not display properly in a standard ISO report suite. Some of these non-supported characters may cause display problems such as line breaks, odd characters, or even truncation of the value passed to SiteCatalyst.

If the data you are passing to SiteCatalyst contains any characters not in the ISO-8859-1 character set, you should use a multi-byte report suite. Contact your Implementation Consultant or Omniture Client Care to make the change. A report suite can be changed from standard to multi-byte, and vice-versa. However, for data that has already been collected characters above ISO 127 may not display properly once the change is made. The best practice is to determine the needed report suite type when the report suite is created.

### 1.3 Using the charSet Property

The charSet property, which is normally set in the JavaScript file, is used by SiteCatalyst to convert incoming data into UTF-8 for storage and reporting by SiteCatalyst. The charSet property is required when sending data to a multi-byte report suite and should NEVER be used with a standard report suite. Setting the charSet property with a standard ISO report suite can result in variable truncation or unexpected character conversion.

The value of the charSet property should match the web page encoding in the META tag or http header, even though the syntax may differ slightly. Although the META tag may use an alias for the encoding, the value of charSet should use the preferred (or official) name of the encoding. Some of the more common encodings with their preferred name and aliases are listed in the following table.

Preferred Name	Aliases
ISO-8859-1	ISO_8859-1,CP819,latin1
ISO-8859-2	ISO_8859-2,latin2
ISO-8859-5	ISO_8859-5,cyrillic
Big5	Big-5
Shift_JIS	SJIS

Since numerous encodings and aliases exist, contact your Implementation Consultant or Omniture ClientCare to confirm the proper value for charSet if it does not appear in the table above.

If a site has different web encodings on different pages, or a single JavaScript file is used for multiple sites, the charSet property can be set to a default value in the JavaScript file and then reset on specific pages as needed to override the default; for example, s.charSet="UTF-8" or s.charSet="SJIS."

### 1.4 Variable Length Limits

Each SiteCatalyst variable has a defined length limit expressed in bytes. For standard report suites, each character is represented by a single byte; therefore, a variable with a limit of 100 bytes also has a limit of 100 characters. However, multi-byte report suites store data as UTF-8, which expresses each character with one to four bytes of data. This action effectively limits some variables to as little as 25 characters with languages such as Japanese and Chinese that commonly use between two and four bytes per character. The character limit is directly related to the characters being used, which makes a predetermined character limit difficult to determine. For multi-byte report suites, the best practice is to limit SiteCatalyst variables to the specific number of bytes for the variable before passing data to SiteCatalyst.

### 1.5 SiteCatalyst Display Language

The SiteCatalyst interface can be displayed in alternate languages using the Language Menu in the interface. Selecting any option other than English causes SiteCatalyst to display using UTF-8 encoding. Displaying a standard report suite using a setting other than English may cause some data to display improperly.

### 1.6 Character Codes 128-255 (ISO vs. UTF-8)

Characters in the range 1-127 are represented by the same byte sequence (actually a single byte) in ISO-8859-1 and UTF-8. However, the characters in the range 128-255 (including all diacritical characters (accent marks)) are represented by a single byte in ISO-8859-1 and two bytes in UTF-8. The difference becomes apparent when changing the report suite type. For collected data, characters in the 128-255 range that display properly in a standard

report suite will not display properly in a multi-byte report suite. Any of these characters that display properly in a multi-byte report suite will not display properly in a standard report suite. Determining the proper report suite type before collecting data is absolutely critical.

## 1.7 Using the charSet Property

Any non-blank value of the charSet parameter will cause data to be converted into UTF-8 for storage. Any characters in the 128-255 range will be converted to the proper UTF-8 two-byte sequence and stored. These characters will not display properly in a standard report suite. Therefore, the charSet property should never be used with a standard report suite.

Likewise, a blank value of the charSet parameter will bypass the data conversion process, and any characters in the range 128-255 will be stored as a single byte. These characters will not display properly in a multi-byte report suite since the single-byte codes for these characters are not valid UTF-8. Therefore, the charSet parameter should always be used with a multi-byte report suite. Additionally, the proper value should be used with respect to the web page encoding.

## 1.8 Variable Lengths

For a standard report suite, all characters occupy a single byte by definition. When sending data to a standard report suite, all variable length limits expressed in bytes have the same length limit in characters.

For a multi-byte report suite, data is stored at UTF-8. Each character in UTF-8 encoding can occupy one to four bytes of data, which means all SiteCatalyst variables may have their length limit as low as 25 characters. Additionally, the limit on the number of characters is determined by the characters themselves. For example, in UTF-8 you could have a page name consisting of 100 characters "A." However, the character "A" would have a limit of only 50 characters since its character code (192) requires two bytes for storage.

Languages such as French and Spanish frequently make use of diacritical characters. Since each of these characters occupies two bytes of data when stored as UTF-8, variable length limits become an issue. With languages such as Japanese and Chinese, the issue is more profound since each variable can be limited to as little as 25 characters.

Compounding the issue is that if you simply pass a longer variable to SiteCatalyst, the string will be truncated at the byte limit when the data is stored, which has the potential of changing the last character displayed since the database may only contain the entire character byte sequence. For web pages using UTF-8 encoding, you can only use JavaScript to properly limit a variable to a set number of bytes before sending it to SiteCatalyst. However, this technique may not be possible with other encodings such as Big5 or Shift-JIS.

## 1.9 Enabling Multi-Byte Support

To enable multi-byte support the following must be done:

1. The multi-byte pages must use a standard language encoding character set.
2. The SiteCatalyst report suite must be multi-byte enabled.
3. The SiteCatalyst code (charSet) must be set to the correct language identifier for a given language-encoded page.

The JS file must define the charSet variable. (All pageviews and traffic are assumed to be standard 7-bit ASCII unless otherwise specified.) Setting the charSet variable, tells the SiteCatalyst engine what language should be translated into UTF-8. Some language identifiers used in meta-tags or JavaScript variables do not match up with SiteCatalyst's conversion filter. Chapter 2 describes the character sets currently supported by SiteCatalyst.

## 2 Supported Character Sets

There are many other single-byte and multi-byte encodings that are used on the web. Some of the more common additional encodings include the following.

Country	2-char Code	Language	3-char Lang Code	Character Set
Hong Kong	hk	HK Trad Chinese	chi	Big5
Taiwan	tw	TW Trad Chinese	chi	Big5
Korea	kr	Korean	kor	EUC-KR
China	cn	Simp Chinese	chi	GB2312
Africa	aa	English	eng	ISO-8859-1
Africa	aa	French	fre	ISO-8859-1
Argentina	ar	LA Spanish	spa	ISO-8859-1
Australia	au	English	eng	ISO-8859-1
Austria	at	German	ger	ISO-8859-1
Belgium	be	Dutch	dut	ISO-8859-1
Belgium	be	French	fre	ISO-8859-1
Bolivia	bo	LA Spanish	spa	ISO-8859-1
Brazil	br	BR Portuguese	por	ISO-8859-1
Canada	ca	Canadian French	fre	ISO-8859-1
Canada	ca	English	eng	ISO-8859-1
Caribbean	cb	English	eng	ISO-8859-1
Central America	ns	LA Spanish	spa	ISO-8859-1
Chile	cl	LA Spanish	spa	ISO-8859-1
Columbia	co	LA Spanish	spa	ISO-8859-1
Denmark	dk	Danish	dan	ISO-8859-1
Ecuador	ec	LA Spanish	spa	ISO-8859-1
Finland	fi	Finnish	fin	ISO-8859-1

France	fr	French	fre	ISO-8859-1
Germany	de	German	ger	ISO-8859-1
Hong Kong	hk	English	eng	ISO-8859-1
India	in	English	eng	ISO-8859-1
Indonesia	id	English	eng	ISO-8859-1
Ireland	ie	English	eng	ISO-8859-1
Italy	it	Italian	ita	ISO-8859-1
Malaysia	my	English	eng	ISO-8859-1
Mexico	mx	LA Spanish	spa	ISO-8859-1
Middle East	me	English	eng	ISO-8859-1
Netherlands	ni	Dutch	dut	ISO-8859-1
New Zealand	nz	English	eng	ISO-8859-1
Norway	no	Norwegian	nor	ISO-8859-1
Paraguay	py	LA Spanish	spa	ISO-8859-1
Peru	pe	LA Spanish	spa	ISO-8859-1
Philippines	ph	English	eng	ISO-8859-1
Portugal	pt	PT Portuguese	por	ISO-8859-1
Puerto Rico	pr	LA Spanish	spa	ISO-8859-1
Singapore	sg	English	eng	ISO-8859-1
South Africa	za	English	eng	ISO-8859-1
Spain	es	Spanish	spa	ISO-8859-1
Sweden	se	Swedish	swe	ISO-8859-1
Switzerland	ch	French	fre	ISO-8859-1
Switzerland	ch	German	ger	ISO-8859-1
Thailand	th	English	eng	ISO-8859-1
United Kingdom	uk	English	eng	ISO-8859-1

United States	us	English	eng	ISO-8859-1
Uruguay	uy	LA Spanish	spa	ISO-8859-1
Venezuela	ve	LA Spanish	spa	ISO-8859-1
Vietnam	vn	English	eng	ISO-8859-1
Estonia	ee	Estonian	est	ISO-8859-10
Croatia	hr	Croatian	cro	ISO-8859-2
Czech Republic	cz	Czech	cze	ISO-8859-2
Hungary	hu	Hungarian	hun	ISO-8859-2
Poland	pl	Polish	pol	ISO-8859-2
Romania	ro	Romanian	rom	ISO-8859-2
Slovak Republic	sk	Slovak	slk	ISO-8859-2
Slovenia	si	Slovenian	slv	ISO-8859-2
Lithuania	lt	Lithuanian	lit	ISO-8859-4
Bulgaria	bg	Bulgarian	bul	ISO-8859-5
Ukraine	ua	Russian	ukr	Windows-1257
Russian Federation	ru	Russian	rus	Windows-1257
Greece	gr	Greek	gre	Windows-1257
Turkey	tr	Turkish	tur	Windows-1257
Israel	il	Hebrew	heb	Windows-1257
Latvia	lv	Latvian	lat	Windows-1257
Japan	jp	Japanese	jpn	SJIS



CALL 1.877.722.7088  
1.801.722.0139

[www.omniture.com](http://www.omniture.com)  
[info@omniture.com](mailto:info@omniture.com)

550 East Timpanogos Circle  
Orem, Utah 84097

