Summary of Unicode Encoding Rules

Character encoding

- ASCII characters (in the range 0-127) are encoded as a single byte.
- Greek, Hebrew, Arabic and most accented European characters are encoded as two bytes;
- Other characters are encoded as three bytes;
- The individual characters are encoded according to the following rules.

Single byte encoding

Characters in the range 'u+0000' to 'u+007f' are encoded as a single byte.

Table 1: UTF-8 Single Byte Encoding

byte 0	
0	bits 0-6

Two byte encoding

Characters in the range 'u+0080' to 'u+07ff' are encoded as two bytes.

Table 2: Two byte encoding

byte 0				byte 1		
1	1	0	bits 6-10	1	0	bits 0-5

Three byte encoding

Characters in the range 'u+0800' to 'u+ffff' are encoded as three bytes:

Table 3: UTF-8 Three Byte Encoding

byte 0					byte 1			byte 2		
1	1	1	0	bits 12-15	1	0	bits 6-11	1	0	bits 0-5

Notes on encoding rules

The first bits of each byte indicate the role of the byte. A zero bit terminates this role information. Thus possible byte values are:

Table 4: UTF-8 Encoding Rules

Bits	Byte value	Role
0???????	000-127	Single byte encoding of a character
10??????	128-191	Continuation of a multi-byte encoding

110?????	192-223	First byte of a two byte character encoding
1110????	224-239	First byte of a three byte character encoding
1111???	240-255	Invalid

Example encoding

Table 5: UTF-8 Encoding Example

Character	**S**	**C**	**T**	®		③		
Unicode	0053	0043	0054	00AE		2462		
Bytes	01010011	01000011	01010100	11000010	10101110	11101111	10111111	10111111