**Unicode**

Character representation system used to map (encode) all characters, present and future

Currently 128,000 characters represented in Unicode, possible to expand this to 1.2M characters

**UTF** (Unicode *Transformation* Format.)

UTF8, UTF16, UTF32

UTF8 is a (variable length) encoding scheme that encodes Unicode character using either 1,2,3, -> 6 bytes

UTF-8 Example:

|  |  |  |
| --- | --- | --- |
| **Unicode** Value | **UTF8** Encoding | #Bytes |
| U-00000000 – U-0000007F: | 0*xxxxxxx* | 1 |
| U-00000080 – U-000007FF: | 110*xxxxx* 10*xxxxxx* | 2 |
| **U-00000800 – U-0000FFFF:** | 1110*xxxx* 10*xxxxxx* 10*xxxxxx* | 3 |
| U-00010000 – U-001FFFFF: | 11110*xxx* 10*xxxxxx* 10*xxxxxx* 10*xxxxxx* | 4 |
| U-00200000 – U-03FFFFFF: | 111110*xx* 10*xxxxxx* 10*xxxxxx* 10*xxxxxx* 10*xxxxxx* | 5 |
| U-04000000 – U-7FFFFFFF: | 1111110*x* 10*xxxxxx* 10*xxxxxx* 10*xxxxxx* 10*xxxxxx* 10*xxxxxx* | 6 |

* **One byte** : looks *exactly the same in UTF-8 as ASCII,*
* **Two bytes**: Most European and Middle East letters, extended Latin letters (with tilde, acute, grave and other accents), Cyrillic, Greek, Armenian, Hebrew, Arabic, Syriac, and others.
* **Three or Four bytes** : Korean, Chinese, and Japanese ideographs, [CJK]

**Examples**:

‘H’ Unicode Representation is U+**0048,** requires 1 bytes to transform it from Unicode to utf8.

|  |  |
| --- | --- |
| U-00000000 – U-0000007F: | 0011 0000 |

Becomes UTF-8 **0x48**

So Unicode and ASCII same for first 128 characters

Euro ‘€’ U+**20AC,** requires three bytes to transform it from Unicode to utf8.

|  |  |
| --- | --- |
| U-00000800 – U-0000FFFF: | 1110*xxxx* 10*xxxxxx* 10*xxxxxx* |

Becomes:

|  |  |
| --- | --- |
| U-00000800 – U-0000FFFF: | 1110***0010*** 10***000010*** 10***101100*** |
| = UTF-8 **0xE282AC** |  |

U/**20AC** equivalent to **UTF8 0xE282AC**

Windows provides APIs based on UTF-16LE for reading from and writing to the console. As of MySQL 5.6.2, the [**mysql**](https://dev.mysql.com/doc/refman/5.6/en/mysql.html) client for Windows is able to use these APIs.

1. Open a console window.
2. Go to the console window properties, select the font tab, and choose Lucida Console
3. Execute [**mysql.exe**](https://dev.mysql.com/doc/refman/5.6/en/mysql.html) with the [--default-character-set=utf8](https://dev.mysql.com/doc/refman/5.6/en/mysql-command-options.html#option_mysql_default-character-set) option.

Alternatively modify my.ini config file:

[mysql]

default-character-set=utf8

<https://en.wikipedia.org/wiki/List_of_Unicode_characters>

<http://www.ltg.ed.ac.uk/~richard/utf-8.cgi?input=%E2%82%AC&mode=char>

<http://www.joelonsoftware.com/articles/Unicode.html>

<http://www.cl.cam.ac.uk/~mgk25/unicode.html#utf-8>

<http://mysql.rjweb.org/doc.php/charcoll>