The difference between UTF-8 and Unicode?

If asked the question, “What is the difference between UTF-8 and Unicode?”, would you confidently reply with a short and precise answer? In these days of internationalization all developers should be able to do that. I suspect many of us do not understand these concepts as well as we should. If you feel you belong to this group, you should read this ultra short introduction to character sets and encodings.

Actually, comparing UTF-8 and Unicode is like comparing apples and oranges:

UTF-8 is an encoding - Unicode is a character set

A character set is a list of characters with unique numbers (these numbers are sometimes referred to as “code points”). For example, in the Unicode character set, the number for *A* is 41.

An encoding on the other hand, is an algorithm that translates a list of numbers to binary so it can be stored on disk. For example UTF-8 would translate the number sequence 1, 2, 3, 4 like this:

```
00000001 00000010 00000011 00000100
```

Our data is now translated into binary and can now be saved to disk.

All together now

Say an application reads the following from the disk:

```
1101000 1100101 1101100 1101100 1101111
```

The app knows this data represent a Unicode string encoded with UTF-8 and must show this as text to the user. First step, is to convert the binary data to numbers. The app uses the UTF-8 algorithm to decode the data. In this case, the decoder returns this:

```
104 101 108 108 111
```

Since the app knows this is a Unicode string, it can assume each number represents a character. We use the Unicode character set to translate each number to a corresponding character. The resulting string is “hello”.

Conclusion

So when somebody asks you “What is the difference between UTF-8 and Unicode?”, you can now confidently answer short and precise:

UTF-8 and Unicode cannot be compared. UTF-8 is an encoding used to translate binary data into numbers. Unicode is a character set used to translate numbers into characters.

Want more?

- The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About Unicode and Character Sets
- List of Unicode characters
- Unicode Consortium