Unicode Stinks

An essay by Blue-Maned_Hawk

(HTML version | Source | Previous revision | Previous revision source)

Unicode is the single most widely-used standard for encoding text on computer systems. It is significantly better than everything that came before it; instead of a computer needing to juggle eighty different ways to interpret a piece of text, it only needs to deal with one standard that comprehensively contains damn near every writing system in the world. This has significantly improved the computing world, and Unicode is certainly one of the best inventions in recent times.

However, just because Unicode is better than everything else doesn't mean it's good. There are flaws both in the design of the standard and in how it is maintained. It has many redundant or unnecessary characters, it's extremely complicated in how it works, its scope is inconsistent, and it's absolute commitment to backwards-compatibility means that many of the flaws it has can never be fixed. There is so much room for improvement, and in this essay i seek to descibe what a replacement could look like.

A replacement ought to use a fixed-width thirty-two bit encoding; this will ensure that plenty of space is available for new characters (though the most significant bit should be reserved just in case), and any reasonably good compression scheme would counteract the space taken up by such large characters. A fixed-width system also simplifies parsing significantly. The characters under this standard ought to be assigned to codepoints hierarchically.

One thing that i think this encoding ought to have is to mark certain characters as invalid—not noncharacters, not reserved or unassigned characters, not private-use characters, but completely and entirely invalid. I think that good candidates for these would be any character where any of the *bytes* (by which i mean sequence of eight bits *aligned to an eight-bit boundary*) is empty, full, or alternating. This would mean that, for example, a binary file couldn't be misinterpreted as a text file, or in an embedded system a jump to a zeroed-out section wouldn't be seen as real characters.

This new encoding ought to take advantage of character composition more than Unicode does—for example, instead of having an analog to Unicode's Mathematical Alphanumeric Symbols block, it could have variation selectors for specifying that a character ought to be rendered in serif or bold or monospace. Variation selectors could also be used in a number of other places—for example, to specify whether a zero should be rendered with a slash or a dot in it.

Whatever group maintains this standard must be okay with breaking backward-compatibility from time to time. The preservation of back-compat may be useful, but it cannot be done ad infinitum. In all things, historical cruft () will build up, and eventually an effort must be made to combat this for the betterment of the users of the system.

I don't have a good way to end this essay.

Should you choose to exercise any of these rights, you must give clear and conspicuous attribution to the original author, and you must not make it seem in any way like the author condones your act of exercising these rights in any way

Should you choose to exercise the second right listed above, you must make this license clearly and conspicuously available along with the original work, and you must clearly and conspicuously make the information necessary to reconstruct the work available along with the work.

Should you choose to exercise the fourth right listed above, you must put any derived works you construct under a license that grants the same rights as this one under the same conditions and with the same restrictions, you must clearly and conspicuously make that license available alongside the work, you must clearly and conspicuously make the information necessary to reconstruct the work available alongside the work, you must clearly and conspicuously describe the changes which have been made from the original work, and you must not make it seem in any way like your derived works are the original work in any way.

This license only applies to the copyright of this work, and does not apply to any other intellectual property rights, including but not limited to patent and trademark rights.

THIS WORK COMES WITH ABSOLUTELY NO WARRANTY OF ANY KIND, IMPLIED OR EXPLICIT. THE AUTHOR DISCLAIMS ANY LIABILITY FOR ANY DAMAGES OF ANY KIND CAUSED DIRECTLY OR INDIRECTLY BY THIS WORK.

ANGZARR