Unicode and Digital Character Encoding

[redacted]
[redacted]
[redacted]
[redacted]

The birth of the internet allowed for unbound communication between countries but with this brought incompatibility with the worlds various writing systems. Unicode quickly became the dominant standard for encoding on the world wide web. This whitepaper sets out to go through the history of how that came to be.

I. Introduction

All computers, most telegraphs and teleprinters use a binary system, where everything is represented with either a 1 or a 0, on or off, long or short. A single digit of binary is referred to today as a 'bit'. These devices opened up a whole new possibilities to communicate with people remotely and with relative ease and speed compared to other forms of remote communication of the time. One of the first of these devices was an electromechanical devise called a telegraph. The telegraph needed a trained operator who was able to encode outgoing messages and decode incoming messages. To make the telegraph more accessible the teleprinter was invented. This device had a keyboard much like a typewriter and the device itself did all the encoding and decoding. No longer did people have to be trained in coding to communicate remotely. These devices were eventually made obsolete by the computer and the internet which is still in heavy use to this day.

Humans are taught their ABC from a very young age and is something they wont ever forget. If electric or electromechanical communication devices can only understand 1s and 0s how do we teach it to understand our ABC? Character encoding can solve this problem by assigning groups of bits to represent certain characters. More and more of the world started using various remote communication devices, these devices did not need to be rewired, retooled or modified to understand a certain country's alphabet as these devices only understood binary. All that was needed was for a country to come up with their own character encoding system to translate those 1s and 0s into their respective alphabet, and that's what country's did.

Every character encoding system is built with the same concepts: code points, code space, code pages and characters. Code points is the measurement of how many bits are within a code space. Either one or multiple code points can be used to represent a single character. Different characters can have differing amounts of code points. Not all characters in a character encoding system need the same sized code points. This differing size allows for flexibility when designing a character encoding system. For example more popular characters can be represented with fewer code points saving on a devices resources. Code space is the total amount of code points a character encoding system has to work with. Devices have limited resources and only a designated amount of bits can be assigned to the code space. While planning out a code page making sure that every character fits within the code space is vital. A code page is the set of characters that a character encoding system supports, what code points represent each character and how much code space they need. Their are two types of characters, printing characters and control characters. Printing characters are quite simply the characters the user sees. These characters are not just letters of an alphabet but they can also be symbols like smiley faces, punctuation or geometric shapes. Control characters are characters the user will never see. They control how characters are printed to the screen. Control characters can be as simple as a character that lets the device know that the line has ended and a new line needs to begin. Or they can be as complex as a character that effects the formatting or style of printing characters.

Some common popular character encoding systems include: Morse code, ASCII and Unicode. Morse code was used on telegraphs and first released in 1844 developed by Samuel Morse. This character encoding system had 36 characters, the 26 letters of the alphabet and numbers '0' to '9'. It has a variable code space with a minimum of 1 code point up to a max of 16 code points. ASCII short for American Standard Code for Information Interchange was a character encoding system developed by the American Standards Association and first used on teleprinters in 1963.[1] It has 128 characters and has a fixed code space of 128 code points. Unicode was first released for computers in 1991 and was developed by the Unicode Consortium.[2] It has close to 150,000 characters and is still being updated with new characters annually. It has a variable code space of 1,114,112 code points at max. These systems vary greatly and demonstrate how flexible character encoding can be given the resources and needs of the users.

II. Problem

The internet was invented back in the 1960s and at that time was used primarily by governments, science laboratories and universities. These various places were all situated in America as at this time the internet was not as global as it is today. With the standardization and improvement of various protocols in the 1980s did the internet start to branch out to other countries like Japan, Australia, New Zealand and countries in Europe.[3] Their use of the internet were much the same as Americas. The internet at this point was not the global World Wide Web that we know today. In the 1990s two very important tools were invented that laid the groundwork to open up the internet to the average consumer on their personal computers and tools that would help create an ever expanding World Wide Web. These two tools were HTTP and HTML invented by Tim Berners-Lee in December 1991.[4] These two things are the backbone of the World Wide Web and are still used today, just in greatly improved forms.

The World Wide Web was introduced to the public consumers in 1991 and over time has grown in an insanely large interconnected network of computers serving and receiving vast amounts of information all around the world. The World Wide Web allows users to view various different forms of media such as photos, text, videos and software from anywhere in the world all from a personal computer just as long as it had internet access. As the World Wide Web grew so did the amount of information being served and received as more people around the world started to get personal computers and the speed at which this information could be exchanged did too. The World Wide Web provided a way for countries to communicate with each other at unprecedented speeds.

This newfound speed at which countries were communicating with each other brought a few new problems, one problem in particular was character encoding. Countries before the World Wide Web each had their own character encoding to suit the needs of their language and the technology they had at the time. But now countries were sharing larger and larger amount of text between each other and the text was getting garbled up. The issue was not the internet messing the text up but the issue lies in the computers using different character encodings so that when they went to print the text they received the computer would not know how to correctly decode the text and would print garbled text it thought was correct. Web browsers first tried to solve this problem by having multiple character encodings to assist the computer, but this solution proved to be not very user friendly as some users would have little clue as to what correct character encoding was needed to display the text correctly. User-friendliness was key to the World Wide Web's success, adoption and growth. A new solution would have to be proposed, character encoding had to be standardized but which character encoding should become the standard when so many character encodings of the time lacked many of the worlds other languages? Unicode with its various formats would be the character encoding system that would end up becoming the global standard for the World Wide Web and the standard for personal computers too.

III. Solution

In 1988 five people: Joe Becker, Lee Collins, Mark Davis, Peter Fenwick and Dave Opstad helped conceptualize and then released a draft proposal for a new character coding called Unicode.[5] The name best represents the goal its creators had. They wanted to create "...a **uni**que, **uni**fied, **uni**versal encoding".[5] Unicode was unique in the way that it didn't have fixed code space. This provides the advantage of popular characters using less code space, gives the encoding the ability to expand to a larger code space if the previous code space got full and provides compatibility with other encodings that each used different size code spaces. Unicodes compatibility with other encodings was a great feature and helped with early adoption of Unicode. People could change their documents over from the encoding they were currently using to Unicode with the benefit of not having to re-encode their documents and the ability to view documents encoded differently without having to have or know what it was encoded in. Users could also convert Unicode documents to older encodings with close to no re-encoding.

The first version of Unicode was released in October 1991 and contained support for 24 of the worlds vast writing systems.[2] Unicode came out just as the World Wide Web was starting to grow. Perfect timing for it too become the dominant character encoding for all the worlds documents to be encoded in. Unicode continues to grow with almost consistent yearly updates adding more and more support for the worlds writing systems with a total of 154 writing systems supported as of Unicodes March 2020 update. As of the 16th of November 2020 a survey of the World Wide Web showed that 95% of the worlds websites used Unicode to encode their websites.[6] Not only is Unicode dominant within the internet but also it has become the default standard for the Microsoft Windows operating systems and most Unix-like operating systems such as Linux and Android. Windows dominates personal computers, Linux dominates servers and Android dominates phones all of which are encoded in Unicode.[7]

Unicode shows no signs of slowing down in being the worlds dominant character encoding as Unicode has next to no competition. Few reasons why someone would use other encodings are for legacy support or private encoding. Unicode provides more than just support for encoding the worlds writing systems, Unicode also has a vast amount of symbols, dingbats and emojis and support for writing systems that are no longer used by the modern world. Unicodes support for writing systems no longer in use provides an easy way for historians to encode and archive old writings in a way that could be view by anyone over the internet. All of Unicode is governed by a consortium who review, propose, develop and implement new additions to the Unicode character encoding with the ambitious goal of providing an ever expanding, ever updating "...**uni**que, **uni**fied, **uni**versal encoding".

References

[1]http://edition.cnn.com/TECH/computing/9907/06/1963.idg/
[2]https://unicode.org/history/publicationdates.html
[3]https://en.wikipedia.org/wiki/History_of_the_Internet#TCP/IP_goes_global_(1980s)
[4]https://www.w3.org/Protocols/HTTP/AsImplemented.html
[5]https://unicode.org/history/unicode88.pdf
[6]https://w3techs.com/technologies/cross/character_encoding/ranking
[7]https://en.wikipedia.org/wiki/Usage_share_of_operating_systems