

Unicode

Unicode is now the standard way that computers encode text and manipulate text so that it be represented in many different global languages. Unicode contains more than 100,000 characters as well as lots of other extensions and documentation that allow it to work so well with the internationalisation and localisation of software. The Unicode Standard features character encodings, reference data files and an encoding methodology together with a set of rules for the normalisation and rendering of international scripts. Unicode began in order to expand the possibilities of the current character encodings of the day. Traditional encodings worked well when Roman characters had to be translated into other local scripts, but were unable to process from one non Roman script into another. This processing of two arbitrary scripts is known as multilingual processing and it has become a crucial feature for 21st century computing. Computer markets are expanding at a rapid rate in non English speaking countries and unicode has been a pretty good solution to the translation problems and localisation of software for these markets. {mosgoogle center} Unicode works by taking a step back from language and represents each character with a unique number - a code point, and leaves the specifics of the actual visual appearance of the character up to other software. Unicode has 1,114,112 of these code points which are normally represented in hexadecimal format. These code points are split into different categories and have different purposes depending on which category plane that they lie on. The Unicode format contains format characters and graphic characters among others. Format characters are not actually seen themselves but can have an effect on the appearance of their neighbour characters. As of Unicode 5.1 there are 141 format characters. Graphic characters in contrast are visible and contain either a visible glyph or a visible space. There are 100,507 graphic Unicode characters as of version 5.1. The represented Unicode characters, known as abstract characters, do not correspond directly to this set of format and graphic characters. Some natural language system characters are represented in Unicode by multiple abstract characters, which exist in sequence and are in turn associated with specific code points within the Unicode framework. There are 75 writing systems or scripts covered in Unicode, although several more are still to be included. Even some historical scripts like Egyptian Hieroglyphics are going to be included in the next Unicode revision. Unicode has become the standard both in stand alone programs and in internet applications. Internal processing within operating systems uses Unicode encoding as do all W3C applications. Unicode within email applications has been a little slow to be picked up, however all of the web based email programs now support it fully. Because Unicode has attempted to become an international standard the main criticisms of it seem to be because it has implemented certain scripts in an inconsistent way or left some out all together. This seems completely natural for a system that is trying to be all things for all people, and the more applications and cultures who embrace Unicode, the more complete it will become.