Additional code space

A science fiction story

by

William Overington

24 April 2004

"I wonder if you could please give me some help with this" asked the trainee archivist.

"I do not understand which characters are being used in this document. The coding is in Unicode with some Private Use Area codes, yet there is something unusual."

"How so?" asks the archivist.

"Well, the Private Use Area characters are in groups of three from the U+F4.., U+F5.., and U+F6.. ranges. Suppose that P, Q, R, S, T, U are all hexadecimal characters in the range 0 to F, then each group of three is of the form U+F4PQ, U+F5RS, U+F6TU with no obvious pattern, though P is usually 2 but sometimes 3."

"Ah, then the coding is possibly using the additional character set: one can never be certain using the Private Use Area, yet a consistent pattern of such groups is a good clue as to the meaning."

"The additional character set? I have studied Unicode a lot yet I have never read anything about the additional character set."

"The additional character set is not Unicode, it is a character set designed to be used as an addition to Unicode so that Private Use Area characters can have an absolute value, so that their meaning can be ascertained by archivists .... such as ourselves. Using your example, U+F4PQ, U+F5RS, U+F6TU means code point A+PQRSTU. So, if you can find a copy of the listings of the additional character set, then you should be able to decipher which characters are intended."

"It seems a lot of coding to use three Unicode characters for each additional character."

"Oh, indeed, yet that is only a portal to the system from 16-bit encoding. With a 32-bit system the encoding is efficient."

"Yet how can I be sure that the encoding is from the additional character set? After all, it is being accessed using Private Use Area codes, so the reference could be to some other character set."

"Well, the answer is that you cannot be sure that the encoding is from the additional character set. It is simply that you have a clue to a possible encoding. It is an encoding to try."

"Well, where should I look?"

"Oh, there is access to the additional character set from the document readers, simply go to options and select the option that causes any group of three characters of the form U+F4PQ, U+F5RS, U+F6TU to be treated as a character from the additional character set."

"What sort of characters are in the additional character set?"

"Characters which have not been encoded in the Unicode character set for whatever reason, some which have been turned down for encoding and some for which no application has been made. There are also code points for various precomposed ligatures and code points for markup items such as codes to indicate the colour of the following text and so on. There are code points for a vector graphics system and a multimedia authoring system. The idea is that anyone can get almost anything encoded in the additional code space."

"What are the modalities of getting a character encoded?"

"One simply asks on a mailing list set up for the purpose. The idea is that the request is debated by the people on the mailing list. Often helpful suggestions are made such as to where in the code space the suggested items are to be encoded and any duplications with existing Unicode or additional code space allocations are mentioned. Usually a consensus emerges within about a week and the characters are added into the code space. The idea is to encode every suggestion unless there is some good reason not to do so. The idea is to provide a facility so that archivists can determine what is intended when someone uses code points which are not part of regular Unicode."

"Are there fonts available for the characters in the additional code space?"

"Well, not all of them need a font, because they are formatting characters, such as indicating that the following text should be coloured red, yet for those characters which are displayed fonts are often provided not for the code points in question but for a Unicode Private Use Area code point. The additional code space documentation often provides a mapping from the additional code space to the Unicode Private Use Area so that a copy of a document can be processed by software so as to produce a document using the ordinary Unicode Private Use Area, yet a document about which one knows the encoding, whereas if one had had that document originally the encoding would not have been known. Not that every use of additional code space uses conversion to the regular Unicode Private Use Area: that is just a technique so that many text processing packages not designed with the additional code space in mind can be used to display the text."

"What happens when the additional code space encoding is formatting or vector graphics or something else which needs software to process it?"

"Well, then one needs to use specialist software which recognizes the code points as needing processing and acts upon them. Sometimes that specialist software will convert to the Unicode Private Use Area before processing."

"Why is that?"

"For ease of processing and because sometimes the encoding was developed using the Unicode Private Use Area and the encoding in additional code space is really just an add-on facility to a

Private Use Area encoding so that documents in that encoding can be archived for long term accessibility."

"Yet surely an ordinary Private Use Area encoding could be specified in a document and the document be available?"

"Indeed, that is often done. Yet deciding which Private Use Area encoding is being used may be the problem for a future researcher trying to understand a document. Using the additional code space does produce a trail which is known about by archivists. Each archivist may not know many, or even any, of the codings used in the additional code space yet once the researcher has been guided that the answer to the problem may be that the encoding is using the additional code space, then the problem may be on the way to being solved, for it is just a matter of finding the coding in the additional code space and deciding whether treating the document as being in that coding makes sense."