Issues Regarding Roman Numerals in UCS

David J. Perry

Draft 1, June 2006

I. The Existing Situation

The following characters that relate to Roman numerals are currently encoded in the Universal Character Set.

- the Latin capital letters I, V, L, X, C, D, and M
- a set of Roman numerals in the Number Forms block, U+2160–217F, both upper and lower case forms; these have compatibility mappings to the Latin capital and lowercase letters
- the ancient numeric forms U+2180 ↀ ROMAN NUMERAL ONE THOUSAND CD, 2181 ↁ ROMAN NUMERAL FIVE THOUSAND, 2182 ↂ ROMAN NUMERAL TEN THOUSAND
- an additional character U+2183 Ↄ ROMAN NUMERAL REVERSED ONE HUNDRED which is "used in combination with C and I to form larger numbers"

II. Historical Background

Most people today equate Roman numerals with Latin capital letters, but this it not their origin. The Romans originally had distinct numeric symbols which, over the course of time, were conflated with letters of the alphabet. This conflation took a long time to complete. The early forms of fifty, ↓, ↓, and ⊥, were used almost until the end of the Republic, and the infinity-shaped sign for 1000 for much longer.

III. The Issues

For modern use, as in copyright dates, the Latin capital letters are clearly the easy and proper way to encode Roman numerals. For classical scholars, especially epigraphers, the issue is more complex. Should the Latin letters always be used, and the various shapes handled as glyph variants through smart font technology? Or is it desirable to encode one or two additional characters?

A. Should the characters 2160–217F be used?

These characters were put into the UCS for compatibility with East Asian standards, and in such a context there is no reason to avoid them. They undoubtedly make Asian typesetting easier, since fonts can have each numeral sized to fit in a square, as East Asian ideographs normally do. For documents utilizing Latin script, however, I think they are better avoided. Western users are

accustomed to using the regular Latin letters as Roman numerals; many do not know about this alternate set of characters; they are inconvenient to enter via keyboard; and their use may cause difficulties in searching.

B. The Numeral Six

Sometimes the shape Ⅽ is found in inscriptions—especially Christian ones— for the numeral six. Cagnat (1898), followed by Gordon (1983) describes this as a ligature of V and I. This makes no sense from a calligraphic point of view, and we do in fact have ancient examples of the V-I ligature that take the expected shape \| need better glyph. A much better explanation is that this form of six is taken from the Greek Stigma, which is used for six.[1] The shape Ⅽ found in Latin inscriptions is quite similar to the shapes ϛ Ϛ found in ancient Greek for the numeral six (not so similar to the shape of Stigma Ϛ shown in the Unicode charts). The fact that this form of six is found mainly in Christian inscriptions—Christianity being an import from the Greek-speaking eastern half of the Roman empire—strengthens the identification with Stigma. If we accept this identification, the question then becomes whether to use the Greek Stigma U+03DA to encode this Roman six, or to propose a new character. In general, Unicode prefers not to mix scripts, and U+03DA clearly belongs to Greek.

C. The Numeral Fifty

As mentioned above, the Roman numeral fifty originally had the forms ↓, ⇓, and ⊥, which were later assimilated to the letter L. There are two ways that this numeral could be handled. Smart fonts might treat the three forms shown above as alternate shapes for the letter L, or a new character ROMAN NUMERAL FIFTY EARLY FORM could be proposed. The former makes for easy searching. However, it does not provide a way to distinguish reliably between L and the earlier forms, which epigraphers want to do; furthermore, the three earlier shapes are not appropriate variants for the letter L in any context except specialist use of Roman numerals. For these reasons the new character should be proposed.

D. The Numeral Five Hundred

Five hundred in origin was D, half of the symbol for 1,000 ↀ. This was easily conflated with the letter D at a relatively early stage. However, to mark the numeral as distinct from the letter, the Romans normally wrote the numeral with a horizontal bar Ꭰ. To encode this distinct shape I suggest using D followed by U+0336, COMBINING LONG STROKE OVERLAY (smart font technology is required to get the stroke placed properly over the D). Another option would be to treat the stroked D as an alternate glyph of regular D. Either of these options is better than using U+0110 Đ LATIN CAPITAL LETTER D WITH STROKE, since users can search for the numeral 500 without needing to know about U+0110 or having to search twice, once for D and once for Đ. Unicode has two other characters with similar appearance: U+00D0 LATIN CAPITAL LETTER ETH Ð and U+0189 LATIN CAPITAL LETTER AFRICAN D, which should never be used for the Roman numeral.

[1] I omit here the complex history of this number, which started out as Digamma, acquired cursive and uncial forms, and was eventually conflated with the Sigma-Tau ligature; see Nick Nicholas's web page for complete discussion.

E. The Numeral One Thousand

Unicode already contains the character U+2180 ↀ ROMAN NUMERAL ONE THOUSAND CD. Here we have a character that can easily distinguish the form used through most of Roman history from the later use of the letter M as a numeral. Smart fonts can treat the many alternative shapes (see Figure 1, which omits the very common forms ∞ and ⋈) as variants of this character.

F. The Numeral One Hundred Thousand

If the note in TUS found with U+2183 "used in combination with C and I to form larger numbers" is followed, 100,000 would be represented as CCCIↃↃↃ (seven separate characters). This is indeed an accurate visual representation of one variant of this numeral (see "13th stylization" in Figure 1). There are two issues here. First, the statement in TUS suggests that the process of adding reversed Cs and Cs to a central I can go on to even higher numbers. Because this method of indicating large numbers is awkward, the Romans developed alternatives: a bar placed over a numeral indicated multiplication by 1,000, and a bar plus vertical lines on either side of a numeral indicated multiplication by 100,000. (For the former, see the lower part of Figure 1.) So numerals higher than 100,000 shown in the CIↃ style are exceedingly rare, if they occur at all. I have never seen any in my research, although I cannot say for certain that there isn't an inscription somewhere that has one. Second, as Figure 1 shows, the numbers 1,000, 5,000, 10,000 and 100,000 had many glyph variants. Smart fonts to support epigraphy need to have a single character for 100,000 which can be used as a base for stylistic alternates, as we already have for 1,000, 5,000 and 10,000.

IV. Possible Solution

The best way to proceed is to propose three new characters for the UCS, ROMAN NUMERAL SIX LATE FORM, ROMAN NUMERAL FIFTY EARLY FORM and ROMAN NUMERAL ONE HUNDRED THOUSAND. Fonts designed to support Latin epigraphy will provide glyph variants using smart font technology.

Bibliography

Ifrah, Georges. *Universal History of Numbers.* NEED PUBLSIHER: 2000. English translation of *Histoire universelle des chiffres*, 1994.

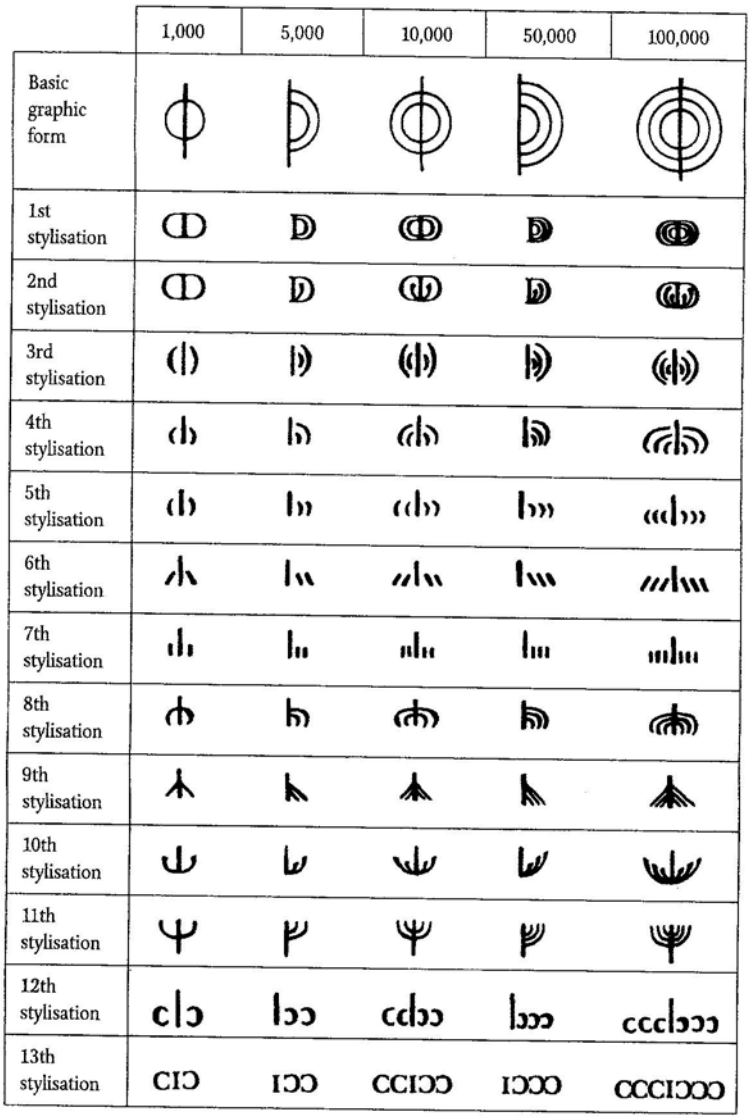

FIG. 16.62.

$\overline{\text{V}}$ = 5,000 = 5 × 1,000	CIL, VIII, 1577
$\overline{\text{X}}$ = 10,000 = 10 × 1,000	CIL, VIII, 98
$\overline{\text{LXXXIII}}$ = 83,000 = 83 × 1,000	CIL, I, 1757

Examples from Latin inscriptions of which the oldest date from the end of the Republic

FIG. 16.63.

IIIIDCCCLXXII	4,872
VDLXVIII	5,568
VDCCCCXVI	5,916
VICCLXIIII	6,264

Examples from a Latin astronomical manuscript of the 11th or 12th centuries CE (Bibliothèque nationale, Paris, Ms. lat. 14069, folio 19)

FIG. 16.64.

Figure 1. From Ifrah 2000, p. 198, showing the variant forms of the larger Roman numerals.