L2/06-256

# Characters with Many Glyph Variants: Some Encoding Issues

David J. Perry
June 28, 2006

The Unicode Technical Committee has raised the question of whether or not it is appropriate to encode as a single codepoint a character that has glyph variants whose shapes are significantly different from the reference glyph and from each other. Here are some thoughts that may help clarify this issue.

It comes down to representing meaning versus appearance. Unicode has chosen to encode characters not glyphs for a number of good reasons. It seems to me that encoding one character in such situations is very much in keeping with this fundamental aspect of Unicode. That's the short answer.

It should be noted that Unicode already contains at least one example of a character with many dissimilar glyph variants, the Greek editorial coronis (U+ 2E0E ⸎). This character has five common variants (☡, ⸏, ℒ, ⸎, ⸎) and several other rare ones; note that the first three are nothing like the reference glyph. The Thesaurus Linguae Graecae, who proposed this character, felt that encoding this one charater and treating the other shapes are glyph variants was appropriate. In fact, it is worth noting that the TLG is moving towards encoding the meaning not the appearnce of their texts. When they began, many years ago, they always tried to encode whatever was on the page in the edition they were using. More recently, they have been using the meaning of the text as the main criterion and not encoding every font change or glyph variant.[1]

Now for the long answer.

Scholars in classics, medieval studies, and related fields have been struggling to find the best way to accommodate their needs within Unicode's character/glyph model. We understand that Unicode will be the basis for storage and manipulation of text on computers for the foreseeable future and we don't want to be left out. But it's not easy. Here is an example.

Epigrapher A wants to email Epigrapher B about a newly discovered inscription which includes the symbol for the Roman sestertius coin in the form Ħ rather than the more common 𐆘. He uses plain text for email, and does not care what font the recipient uses (as long as it has the character in question), what size it is viewed at, what margins it is printed with, or whatever. This is clearly plain text; but he would like to preserve the Ħ shape as found in the inscription.

[1] Information from Richard Peevers of the TLG staff , via email.

So what to do? A reality check indicates that Unicode won't encode five separate sestertius characters because that would fly in the face of the character/glyph model. So our choices are 1) to have one sestertius character or 2) to have no standardized way of representing the sestertius.

There are three advantages to 1).

- it gets the character into Unicode. Otherwise there is no standard for text interchange and every font maker does his or her own thing. Interchange of information is more difficult than it should be, since the character is meaningless if the recipient does not have exactly the same font as the creator of the document.
- it is simple and straightforward; the question "What's the Unicode for sestertius?" has one and only one answer
- it makes searching easy. This is particularly important when dealing with large databases such as the Greek texts produced by the TLG, the Latin texts from PHI, and the Perseus Project. Not every user might know of every glyph variant, and having to do multiple searches is inconvenient and should not be necessary.

Encoding one sestertius character does not address the problem raised in the email example above. There are various ways to address this issue. OpenType or AAT fonts can contain alternate glyphs while preserving the standard Unicode value. Electronic editions can use markup language with entities defined to preserve particular character shapes. Neither of these options is as convenient as we would like, but they do exist.

Now what about a situation in which the various glyphs are less similar to each other than the sestertius glyphs? If Unicode has two, three, or four characters that correspond to the Platonic Idea of one character, we lose the advantages of simplicity and easy searching. If *all* the various shapes are available in Unicode, this might be a worthwhile tradeoff since we can have an accurate representation of the various signs. If, however, only some of the shapes would be available, this represents the worst scenario, since we lose searchability and still do not have all the glyphs as separate codepoints.

The issue of searchability and wide use is crucial here. There are two types of publication, broadly speaking. In the first the editor reproduces the source (inscription or manuscript) exactly as it appears, followed by a normalized presentation that indicates his reading of the text. These editions are mainly of interest to specialists; epigraphers and palaeographers, for instance, want to know exactly what character is used in the original source. Such editions may need to use OpenType fonts with alternate glyphs, specially designed fonts, or markup entities, since not every shape will be available in the UCS.

The second type of publication is the normalized edition, usually aimed at a larger audience. Someone who has studied ancient Greek and can read Homer or the Greek tragedians would have a hard time with many inscriptions in their original form, since the alphabets used varied from place to place and from time to time. So an edition of inscriptions that was meant to aid in the study of Greek history, for example, would present the inscriptions in the standard classical Greek spelling. In such an edition the issue is not epigraphical minutiae, but conveying the meaning of the text so that the reader can apply it to the study of history. In such a situation having one standardized character is helpful, since it is easier for the publisher and the reader and (if the text is in electronic form) searches can be made without difficulty.

Let's consider another example, the Roman centurial sign, which may have the following shapes:

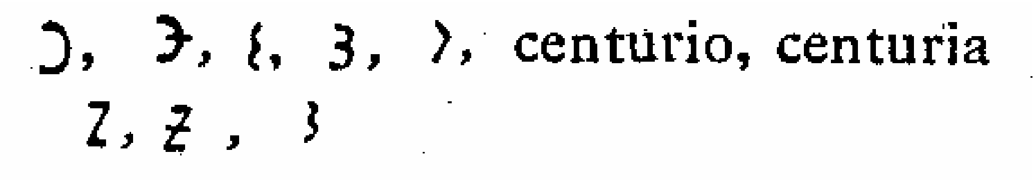
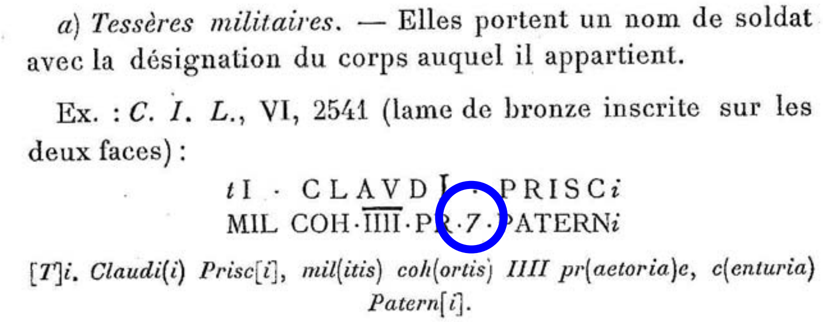


**Figure 1.** From Cagnat p. 445; Capelli p. 516 shows the same signs, except that he does not include the reversed 3 shape.



*a*) *Tessères militaires.* — Elles portent un nom de soldat avec la désignation du corps auquel il appartient.

Ex. : *C. I. L.*, VI, 2541 (lame de bronze inscrite sur les deux faces) :

tI · CLAVDI · PRISCi
MIL COH·$\overline{\text{IIII}}$·PR·7·PATERNi

[*T*]*i. Claudi*(*i*) *Prisc*[*i*], *mil*(*itis*) *coh*(*ortis*) *IIII pr*(*aetoria*)*e*, *c*(*enturia*) *Patern*[*i*].

**Figure 2.** From Cagnat p. 334 showing the 7-shape.

This character illustrates the worst-case scenario mentioned above. If the UTC said "Well, the UCS already contains the reversed C, the angle bracket >, and the letters Z and Ƶ—use one of those" then we would not have simplicity and searchability, nor would we have all the necessary shapes encoded.

So far the argument in favor of one encoded character with several glyph variants has rested on the principles of searchability and simplicity. There is another type of argument that applies to some characters, including the centurial sign. In these cases, a careful palaeographical analysis shows that the various glyphs are more similar than might be apparent at first glance. This provides an additional argument in favor of a single encoding: the disparate glyphs share an underlying identity.

The centurial sign apparently began as a reversed C. There is a tendency in Roman script to change curves to straight line segments, as the samples of Roman cursive preserved on wax tablets, in Pompeian graffiti, and on papyri show.[2] So Ↄ turns into > and 7 and 7. Scribal haste or laziness probably also explains the Z shape as an adaptation of Ↄ, where the horizontals are the top and bottom of Ↄ. Some scribes added a stroke to the ϶ to

[2] This probably has to do with the fact that Romans most often wrote with styli on wax tablets. Drawing curves becomes less convenient when pulling a stylus through a resistant medium.

mark it as an abbreviation, perhaps to distinguish it from the reversed C used for Gaia (never found with a stroke).  From this developed the shapes 3 and ᵌ.  It should be noted that Ↄ and ⁊ are the most common forms.

To summarize:

- the Unicode character/glyph model presents some serious challenges for scholars
- encoding one character, even when glyph variants exhibit significantly different appearance, provides simplicity and searchability and fits into the character/glyph model
- some characters whose variants appear dissimilar at first can be shown to have a common origin or underlying similarity