Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation internationale de normalisation
Международная организация по стандартизации

Doc Type: **Working Group Document**
Title: **Basic principles for the encoding of Sumero-Akkadian Cuneiform**
Source: **Michael Everson and Karljürgen Feuerherm**
Status: **Individual Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2003-05-25**

An informal agenda outline for the second Initiative for Cuneiform Encoding conference (ICE2), to be held 2003-06-5/7 was recently posted to the Cuneiform Discussion List (cuneiform@unicode.org).

Because a number of the suggested agenda items had already been discussed at the first Initiative for Cuneiform Encoding conference (ICE1), and have since then received further consideration by various individuals and groups – for example the Cuneiform Computer Code Project (CCCP; http://members.rogers.com/cuneiform) – and because Michael Everson, a primary representative for Unicode and ISO/IEC JTC1/SC2/WG2 for the encoding of scripts, was not in attendance at that conference, it seemed best to review these questions with him in light of intervening discussions and to record in a formal document our assessment of the prevailing thinking on these matters, in the hope of rendering the ICE2 conference more productive by reducing unnecessary discussion of agreed-upon points.

The questions (as originally posed) and what we believe to be their appropriate answers follow.

1. ***"What exact period will we begin with in our encoding? (We have previously ruled out archaic cuneiform.)"***

It seems agreed by all that to attempt to cover the entire scope of Sumero-Akkadian or Mesopotamian Cuneiform with a single encoding effort would be too great a task for practical purposes. Accordingly, it is our view that the decision as to whether or not to rule out a unification of the archaic script should be deferred pending further study. In any case, from a practical point of view, it is clear that a staged encoding process is called for. If this is so, Question (1) properly becomes "What should be the lower temporal boundary of encoding for Stage One?" Practically speaking, this means choosing a boundary which avoids a minimalist encoding on one hand (reducing the scope to the trivial) and a maximalist encoding, as just rejected, on the other.

During ICE1, the two main candidates proposed for the lower encoding boundary were the Ur III (or Neo-Sumerian) period (*ca*. 2112–2004 BCE by the Middle Chronology – Middle Chronology figures are the better-known and the choice of chronology is of little consequence for encoding in any case) and the Old Babylonian period (*ca*. 2003–1595 BCE). The Old Assyrian period, more or less contemporaneous with the Old Babylonian period but associated with the northern territory of Assyria rather than the southern territory of Babylonia – more properly known as Sumer and Akkad – is under-represented in terms of texts and so is not a suitable candidate. The script of the Ur III period is characterised by crisp and well-defined glyphs, whereas the Old Babylonian period is extensively documented, has been well studied, and is well understood.

In general, all other things being equal, it is our view that we should opt for the earlier period (Ur III) until and unless practical considerations during the actual processing of the material suggest otherwise. (For instance, if we had a lack of qualified cuneiform professionals specializing in Ur III with time to invest, it would likely force us to favour the later, Old Babylonian, period. "Professionals" in this context refers to qualified scholars with extensive theoretical and practical experience, without reference to employment or formal credentials.)

Beginning with an earlier period aids in the identification of splits and reduces the potential of having to disunify many characters at a subsequent stage. In addition, from a representational point of view, if the glyphs chosen for each character are taken from an earlier and more complete repertoire, we ensure more complete pre-merger differentiation in the Unicode glyph table and a minimum of "odd-ball" glyphs. These, typically, will represent additions from the later periods. The inconvenience of under-differentiation of glyphs is much lesser with Ur III or Old Babylonian than if a late period standard such as Neo-Assyrian were chosen, given the high shrinkage of the sign repertoire from early to late periods (there being *ca*. 900 signs in the Ur III/Old Babylonian period, compared with just 600 in the Neo-Assyrian period, and less than 400 in Hittite).

Looking to the future, Stage Two would provisionally encapsulate the Old Akkadian period, while Stage Three would involve treatment of the Archaic Period, should it be decided to unify that as well.

Although we have used the term "period" throughout this section, it is implicit that *all* languages which were clients of the script during those periods should be considered in the event that one or more of these should show peculiarities not evident from the core (Sumero-Akkadian, or mainstream Mesopotamian) tradition.

Questions (2) and (3) should be treated together:

2. ***"How will we treat compound signs, i.e. character sequences made up of two or more signs written one after the other but treated as one grapheme?"***

3. ***"How will we treat complex signs, i.e. signs made up of two or more signs written inside one another and treated as one grapheme?"***

The first question is not well-phrased. It should have been put as follows: "How will we treat compound signs, i.e. linear sequences of two or more signs or wedge-clusters generally treated as a unit?"

Similarly, the second question is better phrased as "How will we treat complex signs, i.e. signs made up of a primary sign with one of more secondary signs written within it or in such proximity to it that the whole is generally treated as a unit?"

The decision reached at ICE1 on these matters was that complex signs (or "inscribed signs", as they are sometimes called) would be encoded as single characters, notwithstanding the possibility of identifying the individual sign elements making up the composition of the glyph, while sequential compounds would be encoded as a sequence of characters, one per constituent element in the sequence. (Although signs which are both compound and complex were not specifically addressed by an ICE1 resolution, their treatment follows in general completely naturally from the other two resolutions.)

No new evidence has to our knowledge been brought forward since ICE1 to require a re-evaluation of these issues, though one clarification should be made. Since the identity of signs as characters within the system is frequently possible only by means of diachronic study of script development in the matter of those signs, it will be necessary, on a sign-by-sign basis, to examine the diachronic development of the sign. Thus, when a sign appears as an (inscribed) complex sign in one place, but as a sequential

compound in another, the sign ought to be encoded as a character sequence; its complex representation could be realized as a ligature via a ZERO-WIDTH JOINER mechanism. The rendering engine will have to resolve such sequences into the appropriate glyph where necessary.

It is unlikely that further round-table discussion of this issue will achieve much in the way of additional encoding principles. In general, we believe it is best to determine whether signs are complex or sequential compounds on an individual basis in light of their etymology, with the proviso that one be flexible, since the actual data, rather than some preconceived set of general principles, should inform the decision.

4. "How do we encode gunu signs, i.e. signs with added 'flourishes'?"

Gunû refers to a cluster of additional wedges, similar to "hashing", which differentiate certain signs from otherwise identical signs. In general, one may think of the sign without gunû as the original or base sign, and the gunufied sign as a graphic derivative. Gunufication does not have a consistent semantic interpretation across the set of signs to which it applies.

There are two possibilities when it comes to encoding gunufied signs: either they are simply encoded as signs in their own right, or they are encoded as their corresponding base signs with a COMBINING GUNU modifier.

At present, we are of the opinion that it may be best to choose the former approach, and to avoid the use of COMBINING GUNU. However, the latter approach may have some practical benefits in light of the tendency for scribes in certain periods to substitute base signs for gunufied forms, since this would facilitate software searches of base signs, base signs followed by gunû, and base signs not followed by gunû. Whether such a view is justified cannot be determined, however, without direct study of the repertoire.

5. "How do we treat diri spellings,. i.e. multiple signs treated as one?"

"DIRI" writing, a term derived from the first line of the lexical series diri=SI.A-*ku*=*wa*-*at*-*ru*, refers to the writing of logograms composed of two or more signs, whose reading or meaning cannot be inferred directly from readings or meanings of the component parts. (Example: Akkadian $puzur_4$ 'concealment' is written as a complex sign whose base is KA (Sumerian 'mouth') with inscribed ŠU (Sumerian 'hand') followed by ŠA (which usually stands for the phonemic sequence /ša/.)

The fact that a single reading extends over multiple signs is an issue with respect to the treatment of transliteration, not one for encoding. In principle, there is no reason not to treat DIRI writings in a manner similar to sequential compounds, with one character being encoded for each constituent component.

6. "What sequence will we impose on the signs?"

This is perhaps the last question which should be asked, once the sign repertoire and its encoding has been established, since one or another ordering may be preferable depending upon the final character set. Nevertheless, there is no harm in offering a perspective on this matter now.

There are several possible options. It is unlikely that one of the several familiar Neo-Assyrian wedge orientation-based orders (e.g. Deimel, Borger *Zeichenliste* [1981], and Borger *Mesopotamische Zeichenliste* [forthcoming]) would be suitable, since Neo-Assyrian glyphs offer insufficient coverage for the Unicode glyph table even at Stage One of the encoding. Therefore, one will have either to find a new shape-oriented order based upon Ur III glyphs (or Old Babylonian glyphs as the case may be), or one could opt for a different analysis altogether. Such an analysis would resemble that commonly used to

order Han characters, whose sequence is determined by "radicals", so that related signs would be ordered near to one another in a fashion similar to what is found in Chinese dictionaries. Using a "radical" order instead of a stroke order has the merit that the unification of subsequent stages won't disrupt the organization of the system. Other possibilities exist, but more cannot be decided until the initial character repertoire has been established.

By this we assume that it will be preferable to have the glyphs ordered in the code charts in a harmonious and visually-searchable presentation.

7. "What period will we use for the representative sign glyphs?"

This question also seems somewhat premature. However, it would be most reasonable, from a coverage point of view, to choose glyphs from an earlier, rather than later instance of the script.

Ur III represents the "crispest" sign forms: glyphs are well-formed, and documents are tidily written, so that at this period we can come most closely to speaking of "graphemic distinction" between signs. In our view, this is the most suitable repertoire for choosing the representative glyph set.

As noted above, there will, of course, be a small number of "odd-ball" glyphs in the code table since there will be a certain number of late period sign additions, anomalous language-specific additions for Indo-European, etc. But choosing early-period glyphs is much preferable to choosing a those of a later period, which are not capable of illustrating pre-merger differentiations and which must fall short of full representational capability by 50% or more. Additions at Stages Two (Old Akkadian) and Three (Archaic, if unified) would differ somewhat from the Ur III style, though to a lesser degree.

8. "How will we choose sign names?"

Although the ancient scribes did leave us some lists of their names for the signs, these lists are of late date, are incomplete, and not particularly suited to use for naming a modern encoding.

In the absence of any compelling reason to take an alternate route, we consider it best to continue to name the signs according to their main value as currently practiced in scholarship, e.g. ASH, HAL, MUG, BA, and so on, indexing the names (such as GIR2) as necessary.

9. "What text elements will we encode? – e.g., column, case, and line dividers?"

Case division is a page layout issue paralleling modern tabulation, not a plain text encoding issue. Column and line breaks can be implemented with line-feeds and carriage-returns as in other scripts.

10. "How will we address numeric and metrological signs?"

Where numbers are concerned, some research needs to be done. Are the variant forms for some numbers (e.g. four, which can resemble  ZA or  GAR) significant or arbitrary? If significant, then multiple characters are called for, if not, then only one. The number set should be unified to the extent that actual practice can be represented.

For metrology, the basic elements should be encoded to the extent they do not duplicate other characters.

11. "How do we deal with mergers and splits, in order to support round-tripping between character and transliteration, for example?"

Transliteration is not a primary concern of the encoding; it is a matter of textual interpretation. We are best advised to avoid concentrating on secondary issues such as this and remain focussed on the primary matter before us, which is encoding.

Mergers are no problem. The glyphs for various characters (to cite our long-standing example, say MASH, BAN2 and BAR) will simply be identical in a font representing a post-merger period. In theory, the scholar simply needs to decide what to encode; in practice, this could be handled automatically by means of a suitable input method.

Splits are no problem either. Starting from Ur III, say, it just so happens that at later points in script development certain glyph variants take on a semantic significance of their own (e.g. TA*, which during the Middle Assyrian period is a mere glyphic variant for TA, but which in Neo-Assyrian royal inscriptions has an independent meaning). In the former case, the new character is encoded where the new meaning is implied, and the pre-split character is encoded otherwise.

The creation of new characters at a later stage poses no difficulties whatsoever. They would simply have no representation in earlier periods.

12. "What about ruby (plain text annotation)?"

Ruby exists and is available for general use. It may require special software, however, which could complicate the rendering engine for Cuneiform as it is unlikely to be available except in software specifically supporting Japanese typography. In this case, a viable alternative would be XML mark-up.

Encoding cuneiform means exactly that: encoding cuneiform. It does not imply plain-text-encoding the conveniences of the scholarly apparatus required for multilingual publication. There are numerous mechanisms available to address this need (consider for example the run-of-the-mill word processor and how it handles footnotes), which should be addressed at a later date or in a different arena, if focus is to be maintained.

Summary statement

Most of the questions posed above were addressed over the course of ICE1, and have been discussed a number of times since, with the general conclusions essentially unchanged.

In our opinion, further theoretical discussion of these points is unlikely to accomplish much. What is needed at this point is 1) to collect all relevant and pertinent data; 2) to attempt the application of these principles in the creation of a draft encoding proposal; and 3) to revise them as necessary in light of our findings.

Finally, we believe that this process is unlikely to be completed in time for a proposal to be presented to UTC at the November meeting. Instead, we suggest that we make a start by beginning research and analysis in the areas outlined above (possibly among others) and aim to present the findings of those sub-studies at the November meeting.