IRGN1383

Annex S Discussion: Notes and Comments

Richard S. Cook, UC Berkeley

Individual contribution

IRG Meeting #29, Adobe Systems Inc., San Jose, CA, USA

漢字形音義
Three Domains of Character Variation

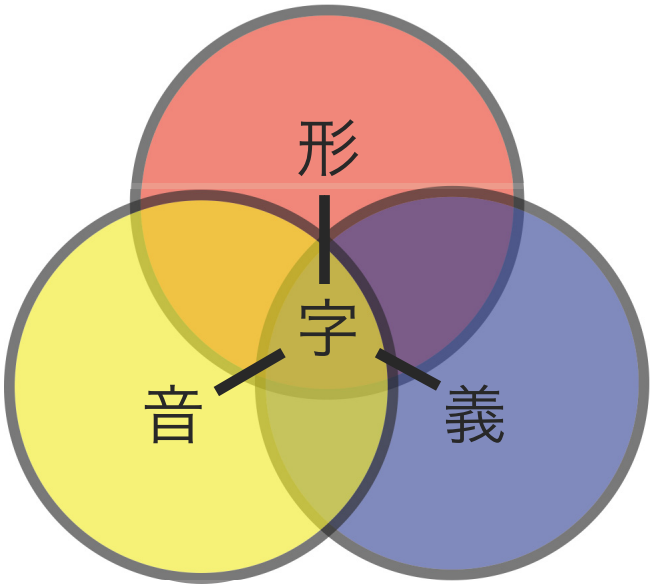
[1] 形體 **Formal**
shape :
representation of real-world thing [physical or mental object];
elements [minimal graphemic distinctive units]:
strokes [type, count, order];
components [assembly of strokes];
coordinates [relative size, position];

[1a] 意義 **Semantic**
meaning :
signification of real-world thing [physical or mental object];
elements [minimal morphemic distinctive units]:

[1b] 聲音 **Phonological**
sound :
signification of real world pronunciation [syllable(s), spoken utterance];
elements [minimal phonemic distinctive units]:

The intersection of these three domains is the domain of character encoding. Items [1, 1a, 1b] in the above list are also in a hierarchy: the *character* (形,字形) is the main unit of writing, encapsulating elements in both *sound* (音) and *meaning* (聲) domains. The semantic domain encompasses both meaning and sound: the character signifies a real world object (not the pronunciation), and simultaneously the character also signifies the pronunciation (which is also a real-world object).

Assertion of identity or difference in each domain is specific to and dependent upon facts and opinions offered in a specific source (national or lexical). Facts and opinions may be present or absent in any given case, in any given domain. Interpretive opinions may have different weight with different interpreters, in different traditions. Two sources of "fact" present in a given domain may differ in evaluating any given case. This is the nature of "opinion": it may be traditional, official, or personal, and is sometimes subjective and difficult to quantify. Though there may be clear "right" and "wrong" within a given tradition, what is a "right" glyph shape in one tradition may be "wrong" in another. Historical study and statistical methods may provide means for assertions which are more confident and more generally acceptable. But historical and statistical methods themselves rest on subjective class judgments (opinions) in the evaluation of edge cases or unfamiliar cases.



In UCS, the representative glyph shape appearing in the code chart is said not to be normative, but only informative. It is called the "representative glyph" because it is chosen from the set of unified variants as representative of the class. Despite the rules of thumb elaborated in Annex S, the class members encoded at a given UCS code point may differ in many ways, including stroke type and component structure, and the individual class members may be distinctive for some purposes. The representative glyph is not the "best example" of the class, but simply one example, one member of the class, standing as one way to represent the class of unified variants. Likewise, Unihan properties for lexical sources are informative, **tracking** opinions contributing to the unification in the encoding process, and guiding "proper" end-user usage after encoding.

Metadata Recommendations

CJK Unified Ideographic Character encoding process metadata recommendations

Formal

- bitmap [image]
- stroke count [stroke type analysis]
- radical assignment(s) [componential analysis]
- IDS → CDL

(Phono-)Semantic

- lexical/print source mapping
- pronunciation(s) per source
- meaning per pronunciation
- variant(s) per source, unified and non-unified

Stroke type assignment is implicit in stroke counting: one cannot count strokes without first identifying individual strokes, and identification involves stroke classification, general or specific (five general types, each with subtypes).

Component analysis is implicit in radical assignment (for internal IRG purposes, or for individual contributor's purpose): one cannot identify an assemblage of strokes as the radical unless strokes are first identified and assembled into higher-level constructs (components).

Rare character are rarely understood! This principle accounts for the many non-distinctive minor variants (encoded and unencoded). Where characters are rare, people do not have clear consensus on form (writing), meaning, or pronunciation. For this reason, the "same" abstract character (by one judgment) may have significantly different concrete manifestations, and it may be impossible to gain consensus on identity judgments.

Characters vs. Components: Encoding rules applicable to independent (free) characters are sometimes not applicable to components. Because components are by definition *parts* of larger constructs (parts of the characters in which they occur), they are smaller than free characters. Fine details differentiating independent characters may be lost when the same character is written smaller for inclusion as part of a larger character. Components occurring at a higher level of a CDL description are naturally larger and "more significant" than characters occurring as components at a lower level of the description (such components tend to be smaller and lack fine detail). Likewise, there may be other contextual deformations of the independent form of a character, when that same character is used in composition (change in stroke type): such component forms are called *combining variants* of the base character (and might be encoded as such, by means of Variation Selectors [VS]). Combining variants, and components in general (including lexical indexing radicals) are problematic in CJK encoding because they function primarily as units in systems of graphological metacharacters: components are used to discuss, analyze and organize other characters (as in dictionaries and in IRG work), but may not be proper "characters" themselves (outside of graphological work analyzing character structure). They lack pronunciations, and are not used to represent morphemes in the spoken language. The set of components is open-ended, since any arbitrarily selected subset of strokes might be assembled for a particular purpose. Such components may also be suitable for standard encoding by means of VS (for example, as a "variant" of one base character in it occurs). If minor component variation (due to information lost in scaling) is not necessarily distinctive, then stroke typing and counting can not provide an infallible method for determining whether two forms are in fact the same abstract character.

Formal considerations must be weighed against lexical source information indicating whether or not two forms are in fact the same abstract character. In cases for which there is authoritative opinion asserting the identity of two forms in semantic and phonological domains, no matter how different the forms may appear, the ideal would dictate their unification. However, if a single lexical source is inconsistent, or if there is no clear consensus, disunification is dictated (if it is deemed appropriate to encode such rare characters). Again, for rare (or locale-specific) characters, there may be insufficient information available for making the determination of identity with an encoded (or candidate) abstract character. In such cases, where information on the meaning of the character is lacking, or where information on the relation of the character to encoded characters is lacking, it may be appropriate to defer encoding until such time as clear usage information becomes available.