L2/07-159

9 May 2007

U-source Database as Versioned Document

John H. Jenkins

As CJK Extension C has been examined by WG2, it's become clear that we need to have a better way to refer to the U-source characters. There are three particular points that underscore this.

One, some UTC members had forgotten that we even had U-source characters in Extension C.

Two, when the time came to supply source information to WG2, we were in some cases unable to point to public documents against which our information could be reviewed.

Three, when Extension C was under review by WG2, WG2 participants were apparently not aware of the fact that the U-source characters are documented in UTC documents when the time came to make corrections in U-source attributions.

Now, Richard Cook and I have been maintaining a FileMaker database with information on all the characters which have been brought to our attention as potential U-source characters. The database includes an index, glyph, source information, and other fields (such as IDS) required by the IRG. CDL data is also included (which is, in fact, used to generate the glyph).

As a note, to avoid confusion, indices are never reused. Even when a character proves to have been already encoded (e.g., UTC00002 = U+221A1) or a determination is made not to encode a character (e.g., UTC00118, which is used in chapter headnotes in Orson Scott Card's novels *Xenocide* and *Children of the Mind*). This means that all later versions of the database are supersets of previous versions.

I have periodically dumped data from this database and submitted it as a UTC/L2 document (e.g., L2/06-364), but there are some problems here. Not only do such documents not include the entire database, but buried as they are in L2 documents, they're not necessarily easy to find even for UTC insiders. For people in general, they may as well not exist.

I recommend that we create a UTR document containing the entire database and relevant fields. This would make it publicly accessible on the Unicode Web site. It would also become possible to make normative references to it, so that in the future, for example, the Unicode Standard and 10646 can point to this document as the U source. Interested parties can then go to the document and obtain further information on character sources.

I will submit to the UTC a complete dump of the database which, with an appropriate header, can be made a draft UTR, if the UTC so decides.