L2/21-150

Specifying Additional Sources for Unihan Database Values

John H. Jenkins

9 August 2021

Summary

There are a number of fields in the Unihan database for which source documents can be specified. (These are currently the `kSemanticVariant`, `kSpecializedSemanticVariant`, and `kZVariant` fields, but it could conceivably be done by the `kDefinition` field in some fashion.) Unfortunately, the formal definitions of these fields do not allow the use of source documents which do not correspond to fields in the Unihan database. Thus the Hong Kong edition of the *Cihai* dictionary can be referenced because it corresponds to the `kCihaiT` field, whereas the mainland Chinese edition cannot.

We suggest here a scheme whereby formal sources can be used without requiring them to correspond to fields in the Unihan database.

Proposal

I suggest that a new table be added to UAX #38 with a list of additional source identifiers. These identifiers would have the same syntax as field identifiers, except they should begin with a different lower-case letter such as “s” (for “source”) or “u” (for “Unihan”). (I prefer “u,” personally.) There is no other data required for these sources other than a full bibliographic reference. For example, we could include in this table:

uSoengMou2003	新商務字典 (*New Commercial Press Dictionary*). (2003) Hong Kong: 商務印書館（香港）有限公司 (Commercial Press (Hong Kong), Ltd.). ISBN 962-07-0140-2

The identifier `uSoengMou2003` would then be available to use by values in the Unihan database.

Any source for which a valid bibliographic reference is available may be included.

To document this, we should add a new section to UAX #38 after §3.8 (“Data Source Identifiers”) to the effect that:

> Fields may include data in their values identifying an authority used as a source. This is currently done by the `kSemanticVariant`, `kSpecializedSemanticVariant`, and `kZVariant` fields. These authorities are indicated by identifiers of two types:
>
> 1) Reference to other fields in the Unihan database. This is done by using the field’s tag (e.g., `kHanYu` indicates that the authority is the *Hanyu Da Zidian*).
>
> 2) Identifiers matching the regular expression `u[A-Z][A-Za-z0-9_]+`. These refer to entries in the table below, containing the identifier and a bibliographic reference to the authority cited.

We may want to move the table to a different section towards the end of the annex in case it gets too long and interrupts the flow of the text.

The regular expressions for the `kSemanticVariant`, `kSpecializedSemanticVariant`, and `kZVariant` will need to be updated to include `[ku][A-Z][A-Za-z0-9_]+` where they currently have `k[A-Za-z0-9]+`.