L2/04-186

Status of the Unihan Database

John H. Jenkins
Apple Computer, Inc.
4 June 2004

Since the last UTC meeting, the following systematic changes have been made to the Unihan database:

The kHKGlyph, kMeyerWempe, and kTang fields have been completed. kFenn is about 85% complete. The kTang field has been revised to use Stimson's romanization (98-A7).

Corrections to the kMandarin field have been made per input from Mark Davis.

The kAlternateMorohashi and kAlternateKangXi fields have been removed (98-A4).

The kSemanticVariant field has been filled out using data generated from other fields (e.g., kMatthews, kMeyerWempe). The syntax has been changed on an experimental basis in the fashion suggested in L2/04-039. (The syntax can be changed back later, if desired.)

We're still waiting for data from Cheung and Bauer (98-A8).

We've added two new internal fields: kFennIndex (which is the actual index of the character in Fenn's dictionary) and kHanyuMandarin (which is the Mandarin readings from the *Hanyu Da Zidian*). We will recommend to the UTC that these be made public when they are fully populated.

Dr. Lu Qin of Hong Kong Polytechnic is generating a full set of Cantonese readings for every ideograph in the BMP. She will supply us a copy of the data for limited use in the Unihan database when she's finished. She is willing for us to put this data in an internal field and to expose it in the on-line database, but she would that the data not be in a field included in the full public text file.