L2/01-417
2 November 2001

Business to Take Up with the IRG

John H. Jenkins
Apple Computer, Inc.
jenkins@apple.com

There are three items involving ideographs which have come up recently and where it would be nice either to have a UTC/L2 decision or blessing for forwarding appropriately to the IRG.

1) Richard Cook of UC Berkeley has noted a large number of errors in the Unihan database field for the *Hanyu Da Zidian*. Our database merely copies the *Hanyu Da Zidian* data directly from IRG sources, so the error originates in their data. Richard is preparing to forward to the IRG on the subject. I've already requested that it be put on the agenda and would like to have UTC approval for moving this forward and trying to get corrected data from the IRG.

2) There has long been a "U-source" indicated in IRG documents and their Super CJK database. This, however, has merely been a place-holder to use for the final Unicode code point assigned a particular Super CJK character. We've requested in the past that the IRG start making a distinction between characters derived from Unicode sources (such as the twelve "compatibility" ideographs noted in the standard) and Unicode code points as targets for IRG unified ideographs. They have noted our desire but not really acted on it.

We've long reserved the right to start submitting ideographs of our own; now may be the time to do it. I've been checking the Unihan repertoire against my Cantonese dictionaries and standard, Western sinological sources such as Karlgren's *Analytic Dictionary of Chinese and Sino-Japanese*, and there are gaps.

In addition, I've received correspondence from Greg Pringle regarding irregular simplifications for some ideographs used to identify certain species of bird with the request that these be considered as candidates for encoding.

I've already set up a FileMaker Pro database on `ftp2.unicode.org` for tracking our candidates for submission and started populating it. It's very small right now, but it can be accessed either by the Web or with a FileMaker Pro 5.5 client. I've also started a group discussion via email on the subject of Unicode submissions to the Extension C

work.

What I'm hoping to get from the UTC (or L2) is:

a) A formal blessing on the work of gathering ideographs to constitute a collection to submit to the IRG for the Extension C work, and

b) Authority to notify the IRG that the US/Unicode *will* be submitting characters for inclusion in Extension C and that it is therefore vital that they begin to distinguish a Unicode scalar value used in encoding a character from their "U-source."

3) This is the tough one.

The correspondence from Greg Pringle and others leads to the whole sticky problem of simplified characters.

Unicode has generally assigned (Chinese) simplified characters their own code points separate from their traditional counterparts; e.g., U+8AAC (說) and U+8BF4 (说). The decision to do so is based on a number of considerations, including the fact that the mapping between simplified and traditional forms is not always one-one and the fact that the IRG G-source distinguishes characters in GB 12345-90 from their simplified counterparts in GB 2812-80.

The problem is that while there are lists published by the PRC of official simplifications, most of these simplifications are obtained by applying a few general principles to specific cases. In particular, there is a set of radicals (such as U+2F94 KANGXI RADICAL SPEECH 言, U+2F99 KANGXI RADICAL SHELL 貝, U+2FA8 KANGXI RADICAL GATE 門 and U+2FC3 KANGXI RADICAL BIRD 鳥) for which simplifications exist (U+2EC8 CJK RADICAL C-SIMPLIFIED SPEECH 讠, U+2EC9 CJK RADICAL C-SIMPLIFIED SHELL 贝, U+2ED4 CJK RADICAL C-SIMPLIFIED GATE 门, and U+2EE6 CJK RADICAL C-SIMPLIFIED BIRD 鸟). The basic technique for simplifying a character containing one of these radicals is to simply substitute the simplified radical.

What this means is that at any time, any publisher of simplified Chinese text may create a new simplified form by merely simplifying the radical. Greg raises the case of U+9D70 鵰, which is a kind of eagle. The "proper" way to write this character in the PRC is to use U+96D5 雕, but Greg has seen U+9D70 written with the simplified "bird" radical instead of the traditional one.

We need to include a general solution to this problem in the standard. I think there are two solutions:

a) Allow the explicit encoding of any form derivable from the existing "traditional Chinese" repertoire in Unihan via the standard simplification rules, perhaps limiting ourselves to only the documented cases.

b) Add language somewhere to §10.1 on the subject of simplified Chinese, to wit:

> There are currently two main varieties of written Chinese, "simplified Chinese" (*jiantizi*) used in most parts of mainland China and Singapore and "traditional Chinese" (*fantizi*), used predominantly in the Hong Kong SAR, Taiwan, and overseas Chinese communities. The process of interconverting between the two is a complex one. This is largely because a single simplified form may correspond to multiple traditional forms, such as U+53F0 台, which is a traditional character in its own right *and* the simplified form for U+6AAF 檯, U+81FA 臺, and U+98B1 颱. Moreover, there are vocabulary differences that have arisen in Mandarin as spoken in Taiwan and Mandarin as spoken in the PRC, so that merely converting the character content of a text from simplified Chinese to the appropriate traditional counterpart is insufficient to change a simplified Chinese document to traditional Chinese. (Note that many Chinese characters are the same in both traditional and simplified Chinese.)
>
> There are two mainland Chinese standards, GB2312-80 and GB12345-90, which are intended to represent simplified and traditional Chinese, respectively. The character repertoires of the two are the same, but the simplified forms occur in GB2312-80 and traditional ones in GB12345-90. These are both part of the IRG G-source, with traditional forms and simplified forms separated where they differ. As a result, the Unicode standard contains a large number of distinct simplifications for characters in the standard, such as U+8AAC 說 and U+8BF4 说.
>
> While there are lists published by the PRC of official simplifications, most of these simplifications are obtained by applying a few general principles to specific cases. In particular, there is a set of radicals (such as U+2F94 KANGXI RADICAL SPEECH 言, U+2F99 KANGXI RADICAL SHELL 貝, U+2FA8 KANGXI RADICAL GATE 門 and U+2FC3 KANGXI RADICAL BIRD 鳥) for which simplifications exist (U+2EC8 CJK RADICAL C-SIMPLIFIED SPEECH 讠, U+2EC9 CJK RADICAL C-SIMPLIFIED SHELL 贝, U+2ED4 CJK RADICAL C-SIMPLIFIED GATE 门, and U+2EE6 CJK RADICAL C-SIMPLIFIED BIRD 鸟). The basic technique for simplifying a character containing one of these radicals is to simply substitute the simplified radical, as in the example above.
>
> The Unicode standard does *not* explicitly encode all simplified forms for traditional Chinese characters. Where the simplified form can be derived by replacing a radical with its simplified form, as a rule, the simplified form should be treated as a font difference. Where the distinction between simplified and traditional forms must be made in plain text, a variant tag should be used. The exceptions to this rule are characters which are marked as having a simplified variant in the Unihan database.

There are additionally some action items for me.

a) We need to go through the Unihan database and proof the simplified variant

information.

b) We need to find all the characters in Unihan which can be simplified by transforming the radical and which do not have explicit simplified counterparts in the standard and add them to our variant tag database.

c) We need to lobby WG2 instruct the IRG not to add any more simplified variants of existing traditional Chinese characters.