22nd International Unicode Conference (IUC22)
Unicode and the Web: Evolution or Revolution?
September 9 - 13, 2002, San Jose, California
http://www.unicode.org/iuc/iuc22/

CJKV Unified Ideographs Extension C

Richard S. COOK
Linguistics Department
University of California, Berkeley
rscook@socrates.berkeley.edu
http://stedt.berkeley.edu/
2002-09-18-10:31

INTRODUCTION

This presentation is concerned with introducing the audience to some of the issues surrounding Ideographic Rapporteur Group (ISO/IEC JTC1/SC2/WG2/IRG) work on "CJK Unified Ideographs Extension C" (Ext C), including the following:

(1)	The IRG methodology constraining glyph submissions for Ext C1 (why more Han characters and which?)
(2)	The method of preparing glyph submissions for the Unicode Technical Committee (UTC)
(3)	IRG member submissions for Ext C1, introducing some of the submitted glyphs, the print sources for the glyph submissions
(4)	The IRG process of submission evaluation
(5)	The impact of submitted glyphs on the "Han Variant" problem (see Cook, IUC-19)
(6)	Plans for Ext C2 UTC submissions

BACKGROUND

As many people already know, The Unicode Standard 3.2 is the best thing ever to happen to the digitization of Chinese texts. The immense work done to produce the CJKV[1] part of this standard, undertaken by the Ideographic Rapporteur Group (IRG)[2], has pushed CJKV computing to higher levels than many had ever thought possible. With the IRG's creation of "Extension B", 42,711 new characters were added to The Unicode Standard, so that it now encodes a total of 70,207 unique "ideographs".[3] The issue is somewhat complicated by things such as "compatibility characters which are not actually compatibility characters". The last totals available to me (provided by Mr. John Jenkins with the advent of Unicode 3.1) are as follows:

Figure 1: Unicode 3.1: Total Unique CJKV Ideographs

27,484	CJKUI, CJKUIA	p. 258 of The Unicode Standard 3.0
27,496	CJKUI, CJKUIA	including 12 compatibility ideographs that are not compatibility ideographs
42,711	CJKUIB	Extension B
70,207	Total number of unique ideographs in Unicode 3.1	

Following completion of Ext. B, the IRG began work to prepare yet more unencoded characters for encoding. This was originally termed "CJK Unified Ideographs Extension C". Preliminary reports from an IRG meeting in Hong Kong indicated that the IRG Rapporteur anticipated submission of some 67,000 candidate ideographs, as these figures (provided to me by Mr. Hideki HIURA) indicate:

[1]Chinese, Japanese, Korean, Vietnamese.

[2]<http://www.cse.cuhk.edu.hk/~irg/>

[3]The term "ideograph" is a technical usage defined in the glossary of the Unicode Standard, a compromise term equivalent to "CJKV character".

Figure 2: Preliminary Ext. C1 Submission Totals

ROK	23000+~20000
TCA	18000
PRC	4570
Japan	~200
Macau	~200
Vietnam	1049
HKSAR	9
DPRK	94

On the basis of these preliminary figures, it was decided to divide submissions into two parts, for Extensions "C1" and "C2". Extension C1 submissions should be those unencoded characters with most immediate relevance to modern usage, while glyphs of less clear status should be reserved for Extension C2 submission.

EXT C1 SUBMISSIONS

At the most recent IRG meeting (IRG-19, held in Macau at the end of April 2002), a total of 26,079 glyphs were submitted by 9 IRG members for inclusion in Ext C1. The breakdown of submissions per member is as follows (sorted by descending number of submissions):

Figure 3: Final Ext. C1 Submission Totals

TCA	**10659**	(Taiwan, ROC)
China	**07650**	(Mainland PRC)
ROK	**04073**	(South Korea)
Vietnam	**02286**	
Japan	**00970**	
UTC	**00271**	(Unicode/US)
DPRK	**00094**	(North Korea)
HK	**00029**	(Hong Kong)
Singapore	**00025**	
Macau	**00022**	
Total	**26079**	

The primary constraint on CJKV submissions is ISO 10646-1 Annex S, which lays out the basic rules determining what the Character Glyph Model means for CJKV. The specific format for glyph submissions required (1) a bitmapped representation of the proposed character; (2) certain tabulated information on each submitted character, including the following:

Figure 4: IRG Submission Format

class:	**Kang Xi**		**Residual Strokes**		**Source**	**Variant**
field:	Virtual Index	Rad. + flag	Count	1st Type	Info & ID	USV
format:	XXXX.YYZ	XXXXY	N	1..5	SSSNNNNN	U1(,U2)
bytes:	1-8	9-13	14-15	16	17-24	25-35

THE UTC SUBMISSIONS

UTC submissions for Ext. C1 were prepared by myself, Mr. Jenkins (Apple Computer), Tom Bishop (Wenlin Software) and Cora Chang (Apple Computer). The process of glyph collection began several years ago with Mr. Bishop's work on the *ABC Dictionary* (University of Hawaii), in which he identified several unencoded simplified characters. Added to this initial batch of candidates for submission were a collection received by Mr. Jenkins from the LDS church in HK. Finally a number of candidate characters were drawn from my own work proofing the Unihan.txt data, and digitizing two large ancient Chinese character lexicons, *Shuowen Jiezi* (c. 121AD)[4], and *Guangyun* (c. 1000AD). Several other simplified candidates for encoding came to our knowledge in emails from Unicode users.

Once the initial candidates for submission had been collected, the hard work began. This included the following:

Figure 5: Steps in Preparing UTC Submissions

(1)	creation of a prototype glyph
(2)	creation of a new record for that glyph in our central "Unihan Additions" database
(3)	entry of relevant data, including glyph prototype (see Figure 4 above)
(4)	checking our candidates against the Unicode CJKV character set

Prototyping of the candidates for encoding was done using undocumented features of a new version of Wenlin software (scheduled for public release in the summer of 2002). Images of each of the candidate glyphs was created using a component-based method, producing images such as the following:

[4]See my IUC-18 paper.

Figure 6: Six Example Prototypes of UTC Ext. C1 Submissions

Once the glyph had been prototyped, it was assigned a UTC number in a record in the "Unihan Additions" database (FileMaker Pro 5). The prototype glyph was placed in a container field of that record, and the accompanying information for that glyph was entered.

Altogether, 312 glyphs were prototyped, though in the checking process (step 4 above) 41 candidates were eliminated as having already been encoded, bringing the final total of UTC submissions to 271.

SUBMISSION REVIEW

The glyphs submitted by the IRG members were pooled and sorted by the IRG Rapporteur and his team, and 4 large PDF's were created, listing the 26078 raw glyph submissions. (This number is one shy of the final total of 26079 glyphs, as 1 additional glyph was voted in after the initial PDF's had been prepared). Several sessions of the Macao meeting were devoted to preliminary evaluation of submissions. The work was divided among the ~40 delegates, and the submission data (see Figure 4 above) underwent the first verification pass. Proofing of the submission data is at present still going on, and it is unclear exactly how much work remains to be done. This will become more clear with the IRG-20 meeting, scheduled for Hanoi in November of 2002.

EXT C SUBMISSIONS AND THE VARIANT QUESTION

In reviewing the glyph submissions for Ext. C1, it appears that the character vs. glyph distinction for CJKV ideographs is still a lively topic of debate. As mentioned in my IUC-19 paper, the distinctions made in ISO/IEC 10646 Annex S do not seem quite up to the task of dealing with the many variant CJKV ideographs. An adequate standard method for quantifying CJKV glyph variation as yet does not exist, though it seems likely that one will in fact be devised on the basis of the Ext B and C work. In my presentation slides I discuss examples of the member glyph submissions, and their relation to encoded glyphs, as well as their relation to the variant question.

UTC SUBMISSIONS FOR EXT. C2

In addition to proofing the Ext C1 submissions, IRG members are now busy collecting and refining submission candidates for Ext C2. The initial mapping work described in my IUC-18 presentation has now progressed to an advanced state, such that I have collected several thousand glyphs which, by the criteria set forth in Annex S are valid candidates for encoding. Many of these glyphs are, however, identified in my mapping tables as variants of encoded glyphs, and should for this reason be treated with a variant selector mechanism rather than being separately encoded. Lacking a mechanism for dealing with such variants, it seems likely that many more variant glyphs will be submitted for encoding in Ext C2. Until the 26079 Ext C1 submissions have been fully digested, it's hard to even begin thinking about Ext C2 submissions. The following is an example of a glyph which might end up in a UTC Ext C2 submission. Whether or not it actually ends up being submitted depends on whether it is somewhere in the Ext. C1 submissions. At the moment of writing, I just don't know for sure. I'll go check, and you do too.

Figure 7: Candidate UTC Ext. C2 Submission

SUMMARY

In summary, it may be said that the IRG's task of evaluating the Extension C1 glyph submissions is an enormous one. The largest problems at present relate to cross-checking the C1 submissions against the enormous encoded characterset. As the encoded characterset grows, such problems only grow with it. The IRG submission and evaluation procedures require much manual human intervention and subjectivity, leaving room for error. As the encoding work continues, guidelines such as those in Annex S must be refined, and standards relating to such things as stroke-count, stroke-type, and component type must codified.

ACKNOWLEDGMENTS

The writing of this paper was supported in part by grants from:

- The National Science Foundation (NSF), Division of Behavioral & Cognitive Sciences, Linguistics, Grant No. BCS-9904950;
- The National Endowment for the Humanities (NEH), Preservation and Access, Grant No. PA-23353-99.

Thanks to John JENKINS and Hideki HIURA for their kind help and suggestions. Thanks also to Tom BISHOP: his work on Wenlin <http://www.wenlin.com/> is now as always an inspiration.