ISO/IEC JTC1/SC2/WG2/IRG N1911
Doc. #: Korea JTC1/SC2 K2180

Korea JTC1/SC2, Committee on Character Codes

Author: KIM, Kyongsok
Date: 2012.11.09.
Status: Individual Expert's contribution
Subject: comments RE: ExtF submission forms of KR, CN, JP, TW, and SAT

1. Background

- The author reviewed ExtF submissions and found that some member bodies followed IRG PnP but some member bodies did not.

- We have IRG PnP N1823 Draft3 which we need to follow in submitting ExtF proposal. The author checked each MB's submission against IRG PnP and summarized the check results below in 3.

- The author also reviewed the file name format specified in IRG PnP. The current explanation seems confusing. As a result, MB's file names have different format. The author suggests to clarify it so that MBs can follow IRG PnP in naming the ExtF file name.

2. IRG PnP N1823 Draft3

- Relevant portions from IRG PnP are quoted below:

d. New CJK Unified Ideographs (Vertical extension). All CJK Unified Ideograph submissions are subject to the following rules:

..

(3) Document Registration: All submission documents should be registered as IRG documents with an IRG document number(IRGN), whose file name should be in the form of:

IRGNnnnn_mmmm[_sss[_ppp]]_submission

where nnnn indicates an IRG document number assigned by the IRG Rapporteur, ***mmmm indicates the member body' s source reference (as listed in 2.2.1.d.(5).a)****, sss can be any member body designated indicator, and* ***ppp indicates the working set or repertoire name (such as Ext. X labelled by "_X").***

(5). The following data for each proposed ideograph must be submitted with **CSV (Comma Separated Value) text format (in UTF-8) or Microsoft Excel format file**:

a) Source reference to indicate the source and the name of the glyph image for tracking. The source reference should begin with a member body abbreviation (G, T, H, M, J, K, KP, MY, U or V)followed by no more than 9 characters and should contain only Latin capital letters, Arabic numbers, and hyphens. The purpose of source references and accepted source references by ISO 10646 are exhaustively listed in Section 23 of ISO 10646. See Annex D for details on information about member body abbreviations.

b) Glyph Image file name. The file name of each glyph image must be the same as the source reference with file extension of .bmp in bitmap format.

c) KangXi Radical Code from 1(U+2F00) to 214(U+2FD5) with an additional 0 or 1 to indicate a traditional character or simplified character, respectively.

d) Stroke Count of the non-radical component (ref. IRGN954AR and IRGN1105).

e) Flag to show whether the ideograph is **traditional (0) or simplified (1)**.

f) Ideographic Description Sequence(IDS) (ref. IRGN1183).

g) Similar Ideographs and Variant Ideographs if available (identified by their code points in the standard in the form of U+xxxxx) or enter "No" if no known variants, leave it empty if not checked.

h) References to evidence documents including document number and page number.

Some sample submissions are listed in Annex G for reference.

Member body abbreviations in this document correspond to the source standard categories in ISO/IEC 10646 Section 23 except MY.

3. Analysis of each MB's submission form

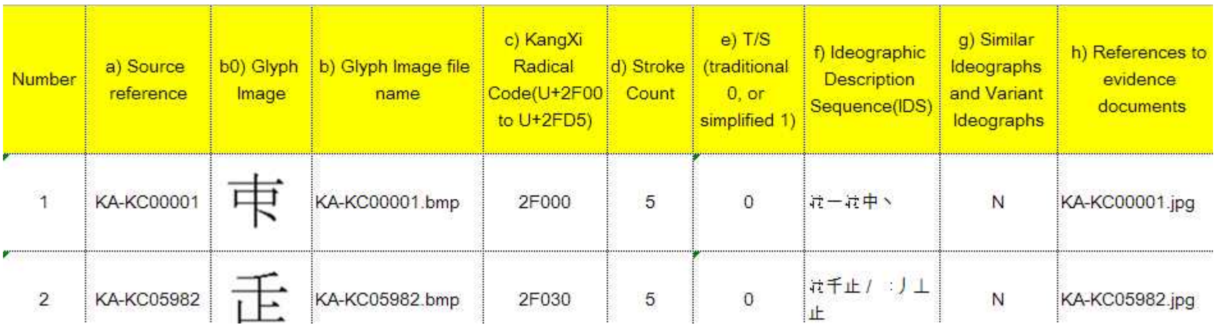
- The analysis is mainly based on column headings and column values; however, a thorough analysis was not done. There could be errors. Comments are welcome.

3.1 KR (ROK)

1) ROK tried to follow IRG PnP as precisely as possible. However, item h) was confusing as explained below.
 - items a) ~ h) are exactly as explained in PnP.

2) ROK has one additional item b0) which is not specified in PnP.
 Rationale: If we do not have glyph image itself, it is hard to check CJK chars.

 - Suggestion: how about adding one NEW item: "glyph image"? (item b0) below)



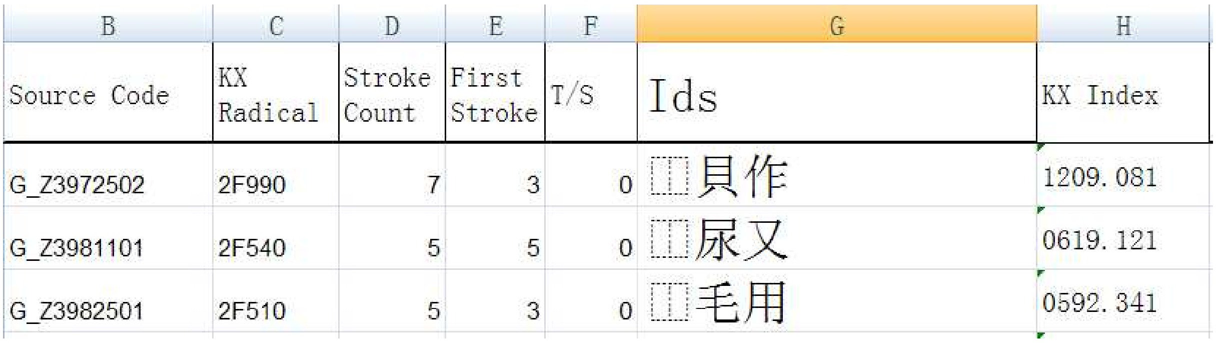
Number	a) Source reference	b0) Glyph Image	b) Glyph Image file name	c) KangXi Radical Code(U+2F00 to U+2FD5)	d) Stroke Count	e) T/S (traditional 0, or simplified 1)	f) Ideographic Description Sequence(IDS)	g) Similar Ideographs and Variant Ideographs	h) References to evidence documents
1	KA-KC00001		KA-KC00001.bmp	2F000	5	0	⿱一⿱中丶	N	KA-KC00001.jpg
2	KA-KC05982		KA-KC05982.bmp	2F030	5	0	⿱千止 / :丿丄 止	N	KA-KC05982.jpg

3) RE: item h)
 ?? what is meant by "document number"?
 If we list evidences and number each of them, then the document number makes sense. If that's what is meant by PnP, then we need to specify cleary so that MBs can easily follow the format.

4) ROK: Currently the value in column h) is not complete. ROK will modify values in column h).

5) Summary: no items are missing; item h) need to be modified later.

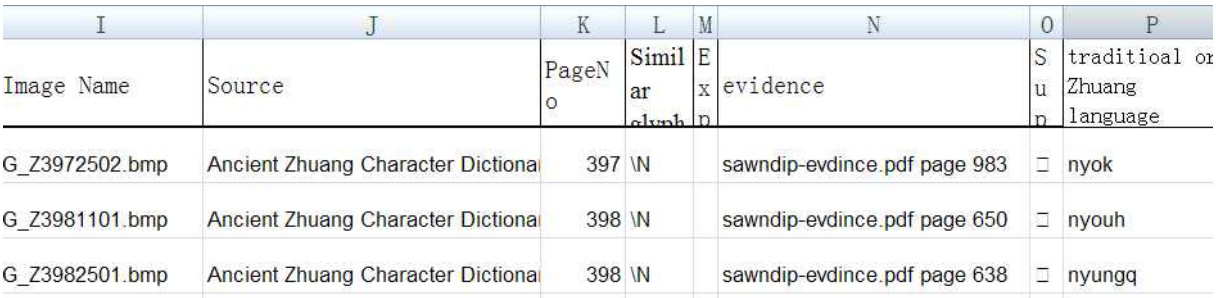
3.2 CN (China)



B	C	D	E	F	G	H
Source Code	KX Radical	Stroke Count	First Stroke	T/S	Ids	KX Index
G_Z3972502	2F990	7	3	0	⿰貝作	1209.081
G_Z3981101	2F540	5	5	0	⿰尿又	0619.121
G_Z3982501	2F510	5	3	0	⿰毛用	0592.341

1) Too many items ^ - ^. IRG PnP askes for only 8 items.
2) Analysis of each column:

B) source code ==> source reference, NOT source code; a) (OK)
C) KX Radical ==> c) (OK)
D) Stroke Count ==> d) (OK)
E) First Stroke ==> NOT needed ==> TO BE DELETED (or HIDED)
F) T/S ==> e) (OK)
G) IDS ==> f) (OK)
H) KX index ==> NOT needed ==> TO BE DELETED (or HIDED)

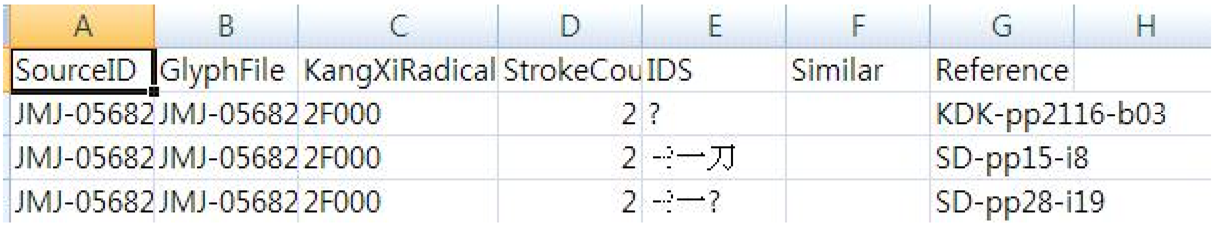


I	J	K	L	M	N	O	P
Image Name	Source	PageNo	Similar glyph	Exp	evidence	Sup	traditioal or Zhuang language
G_Z3972502.bmp	Ancient Zhuang Character Dictiona	397	\N		sawndip-evdince.pdf page 983	□	nyok
G_Z3981101.bmp	Ancient Zhuang Character Dictiona	398	\N		sawndip-evdince.pdf page 650	□	nyouh
G_Z3982501.bmp	Ancient Zhuang Character Dictiona	398	\N		sawndip-evdince.pdf page 638	□	nyungq

I) image name ==> b) (OK)
J) source ==> part of h) ref. to evi. doc (??)
K) page No ==> part of h) ref. to evi. doc (??)
L) similar ==> g) (OK)
M) expl ==> to be DELETED (or HIDED) OR to be MOVED at the end of row
N) evidence ==> ?? part of h) ?? difference between J)+K) and N) ??
O) supp ==> to be MOVED at the end of row
N) trad ... ==> already at the end of row

3) summary: no items are missing;
- need to rearrange items a) to h) according to the order of PnP.
- additional items need to be moved to the end of a row.
- If item h) is divided into several subitems, subitems could be labeled as h1), h2), h3) etc.

3.3 JP



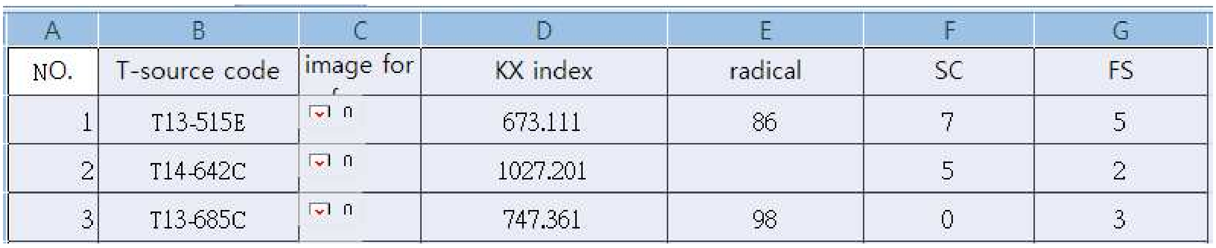
	A	B	C	D	E	F	G	H
	SourceID	GlyphFile	KangXiRadical	StrokeCou	IDS	Similar	Reference	
	JMJ-05682	JMJ-05682	2F000	2	?		KDK-pp2116-b03	
	JMJ-05682	JMJ-05682	2F000	2	-⁝一刀		SD-pp15-i8	
	JMJ-05682	JMJ-05682	2F000	2	-⁝一?		SD-pp28-i19	

1) Analysis of each column:

A) Source~~ID~~ ==> a) source reference, NOT sourceID (OK)
B) GlyphFIle ==> b) (OK)
C) KangXiRadical ==> c) (OK)
D) StrokeCount ==> d) (OK)
E) IDS ==> f) (OK)
F) similar ==> (OK)
G) Reference ==> (OK)

2) **summary: item e) trad/simp: MISSING**

3.4 TW (TCA)



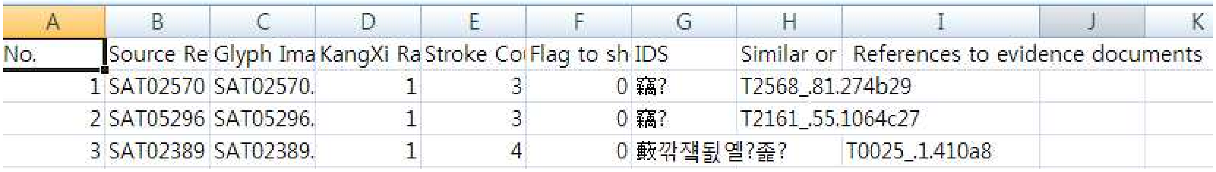
A	B	C	D	E	F	G
NO.	T-source code	image for	KX index	radical	SC	FS
1	T13-515E		673.111	86	7	5
2	T14-642C		1027.201		5	2
3	T13-685C		747.361	98	0	3

1) Analysis of each column:

B) T-source ~~code~~ ==> a) source reference, NOT source code (OK)
C) image for ==> b0) glyph image (see KR section for this item)
D) KX index ==> KX indexe is NOT needed ==> to be DELETED (or HIDED)
E) radical ==> NOT radical no, but radical code pos. is needed
==> to be MODIFIED
F) SC ==> d) (OK)
G) FS ==> NOT needed ==> TO BE DELETED (or HIDED)

2) summary: items b), c), e), f), g) and h) are MISSING

3.5 SAT/JP



A	B	C	D	E	F	G	H	I	J	K
No.	Source Re	Glyph Ima	KangXi Ra	Stroke Co	Flag to sh	IDS	Similar or	References to evidence documents		
1	SAT02570	SAT02570.	1	3	0	竊?	T2568_.81.274b29			
2	SAT05296	SAT05296.	1	3	0	竊?	T2161_.55.1064c27			
3	SAT02389	SAT02389.	1	4	0	敷깍잴됬옐?좗?		T0025_.1.410a8		

1) Analysis of each column:

B) SourceRef ==> a) (OK). ?? does not begin with 'J'
C) GlyphIma ==> b) (OK)
D) KangXiRadical ==> c) NOT radical no, but code pos. of radical needed;
=> to be MODIFIED
E) StrokeCount ==> d) (OK)
F) flag trad/simp ==> e) (OK)
G) IDS ==> f) (OK)
H) similar ==> g) (OK)
I) Reference ==> h) (OK)

2) summary - item c) is MISSING (or to be MODIFIED)

4. RE: submission file name format

IRGNnnnn_mmmm[_sss[_ppp]]_submission where

nnnn indicates an IRG document number assigned by the IRG Rapporteur,

?? mmmm indicates the member body’ s source reference (as listed in 2.2.1.d.(5).a) --> hard to understand. what value should be used?

sss can be any member body designated indicator, and

***ppp indicates the working set or repertoire name (such as Ext. X labelled by** “_X”).*

4.1 Problem #1

- The author tried to follow PnP.
- However, it is not possible to assign source reference to mmmm since there are many different source references in the submission.
- mmmm could be "MB" abbreviation (*G, T, H, M, J, K, KP, MY, U or V)?*
  If that is the case, we need to correct IRG PnP.

4.2 Problem #2

- Since we need to submit not just one but a couple of documents, there must be some mechanism to distinguish them.

- for example, the following file names seem OK?

IRGN1887_K_excel_F_submission.xlsx (for submission form: excel/csv)

IRGN1887_K_sumry_F_submission.doc (for submission summary form)

IRGN1887_K_BMP_F_submission.zip (for BMP files)

IRGN1887_K_listevi_F_submission.doc (for a list of evidences?)
- or a list of evidences could be appended at submission summary form ?

IRGN1887_K_evidences_F_submission.zip (evidences files?)

- How about changing the format so that each MB could put its own indicator at the end of a file name?

  IRGNnnnn_mmmm[_sss[_ppp]]_submission
-->
  IRGNnnnn_mmmm_ppp_submission_sss

- New file names following the modified naming scheme are shown below:

IRGN1887_K_F_submission_sumry_20121020.doc (for submission summary form)

IRGN1887_K_F_submission_BMP_20121020.zip (for BMP files)

IRGN1887_K_F_submission_listevi_20121020.doc (for a list of evidences?)
  - or a list of evidences could be appended at submission summary form ?

IRGN1887_K_F_submission_evidences_20121020.zip (evidences files?)

- summary: Each member body seems to have its own file naming scheme ^ - ^
  We need to have consistent file naming scheme for efficient review.

* There could be errors. Comments are welcome.

* * *

Universal Multiple-Octet Coded Character Set
UCS

ISO/IEC JTC1/SC2/WG2 IRG N1911Comments
Date: 2012-11-11

Source:	John Knightley
Meeting:	IRG#39, Hanoi, Vietnam
Title:	About comments and Reports on Ext F submissions
Status:	Individual submission
Actions required:	For discussion
Distribution:	IRG
Medium:	Electronic

A comparison of submissions for Ext F shows considerable differences between them, furthermore in some cases in their present form they fall short of the requirements set forth in IRG P&P.

1) Considering IRGN 1899 "Preliminary Report on Ext. F submission by China, Japan, Korea, SAT and TCA"

Of the five submissions received, one submission TCA IRGN 1885 did not include IDS and so could not be included in the report IRGN 1899, and for the three type of possible duplications 221 out of 301 came from one submission SAT.

IRGN 1899	Maybe Encoded	Maybe Ext E	Maybe Same Source	Total
CHINA	10/1345=0.7%	4/1345=0.3%	0/1345=0.0%	14/1345=0.1%
JAP	36/1834=2.0%	8/1834=0.4%	8/1834=0.4%	52/1834=2.8%
KOREA	20/1973=1.0%	4/1973=0.2%	1/1973=0.1%	25/1973=1.3%
SAT	186/3515=5.3%	12/3415=0.4%	23/3415=0.7%	221/3415=6.5%
TOT	252	26	23	301

The above plus the facts that in IRGN1883Sat_extf_121018.csv 100 IDS contain "？" and 321 proposed characters have "Similar or Variant Ideographs" in UCS make it clear that is is very probable that more than 5% of the submissions are duplicates that should not have been submitted, and certain that if accepted in it's present form, then a disproportionate amount of time would be spent on problems to do with the SAT proposal.

As IRGN P&P 2. says "(2). **Pre-submission Unification Checking**: A member body should be EXTREMELY CAREFUL ***not to submit CJK Unified Ideographs that are already standardized or previously discussed*** and recorded at IRG meetings. " This has clearly not been done to the same degree for the SAT proposal as for others.

2) Considering IRGN 1911 "Comments RE: ExtF submission forms of KR, CN, JP, TW, and SAT"

This points out differences in naming and where submissions have missed out certain items. There would seem to be two items that are particularly serious problems.

One is the number of items missing for the TCA submission, such that have already disrupted the normal flow of processing submissions in that it was not possible for a report to be made based on the IDS. It is important the submissions be sufficiently complete so as not to impede efficient processing of submissions

The other IRGN 1911 asks the question as to why the SAT submission does not begin with J . The IRGN P&P states clearly in many places that submissions are to be made by member bodies, in for example:-

"2.2.1. Basic Rules on Submission

A member body may submit the following to the IRG along with its repertoire."

The fact that characters are first submitted to an individual member body which then compares the characters to UCS, IRG docs, and other characters received by that individual will invariably lead to a removal of duplicates, and ensures that an extra level of checking and preparation takes place before the characters are submitted increasing the quality of the combined submission made by the individual member and to the IRG thus greatly reducing the work load placed on all members of the IRG. In the event that it is not possible for an individual body to process all characters suggested to it in time for a particular extension, or when the combined total is more than the permitted number for an extension then the individual member is required by the IRGN P&P to make a submission that are only part of the total characters suggested to it.