Ext.-C ad-hoc group Report

An ad-hoc group for preliminary discussion on “extension-C” was held at #17 IRG meeting.

Meeting was from afternoon 18th June to morning 21st June.

Following IRG members were in the ad-hoc discussion:

Mr. Zhang, Guoqiang	China.	Ms Chan, Teresa	HK
Mr. Pan, Cheng-Wu	TCA	Mr. Kang, Myong Do	DPRK
Ms Wei, Lin-Mei	TCA	Mr. Ko, Song Hak	DPRK
Mr. Ngo, Tung Viet	Vietnam	Mr. Lee Joon Suk	ROK
Mr. Wu,Lieh-Neng	TCA	Mr. Lee Keon Sik	ROK
Mr. Chau, C.K. Clement	Macau	Mr. Hideki Hiura	Unicode
Mr. Tou, C.M. Joe	Macau	Mr. TK. Sato	Japan (lead)
Mr. Chan K.K. Arnie	Macau		

1. Extension-C proposal status review.

The ad-hoc group reviewed an up-dated status of country proposals.

As of 2001-06-21, status is as follow.

DPRK	N798	94 chs, old proposal is withdrawn.
Vietnam	N804	1,049 chs (under unification rule)
Hong Kong	N811	9chs after cleaning of an application by unification rule
China	N818	4,570 chs, clean new chs., old proposal with drawn
TCA	N819, N819+	18 K chs, mostly for name and address, possibility of some compatibility ideographs.

Singapore	N828	25 chs (proposal was originqlly submitted to WG2)
ROK	N823	23,385 chs (mostly found in Korea tripitaka)
		Continuing a review, about 20K chs to be proposed at #18 meeting
Japan	(under investigation)	a couple of hundreds?
Macau	(in process)	around 200 chs not defined in BIG-5. not yet checked with SCJK, The proposal will be submitted after the completion of the review.

More than 65K in total.

2. Characteristics of the proposed CJK ideographs ext.-C

The proposed characters have following characteristics.

a. One key source of the characters under proposal is personal name, company name, address

b. Another key source of the characters is classical documents (for digitalization). ROK expressed its strong feel on the needs of those characters on UCS. There were no objection on this view among the ad-hoc members and they support the opinion.

c. Also, there are still “conventional newly proposed” mass user usage characters.

d. Most proposed characters are low frequency usage, but some countries are still early stage of collecting the necessary characters. This means that:

There is a possibility of very high priority characters within a proposal from such countries. And,

There will be a proposal for low frequency usage characters from those countries in later.

e. Also, most of the country expressed the possibility of the future addition (again), however, except for ROK, DPRK and Japan, the possibility of the “more” is a matter of long range. (not in 2-3 years)

f. There are many characters that are cognate with the CJK ideographs which are already coded

within UCS. But the glyph shapes are different (out side existing unification rule). There is a need of discussion on those characters prior to character by character review.

g. And also, there is still significant number of non-cognate characters in proposals.

h. Most of small number proposers are following the unification rule and not seeing a request beyond current unification framework.

i. Most of the proposals are reasonably stable per national body say (possible minor addition while review process). But ROK, DPRK, Japan, and Macau will bring in a significant amount of characters soon.

j. Still there is a possibility of addition of large amount in future (such as ext.-D project)

3. Information (at what quality) to be submitted as a part of proposal

The ext-C ad-hoc group discussed necessary information to be submitted as a part of national Ext-C proposal.

In addition to the information the IRG used to request, the group agreed that an addition of following information along with the proposed characters might help a productivity and quality of developing process of extension-C by IRG

a. Alternative radical: In some case, there will be a disagreement of radical selection between the proposals. It would be better to provide all possible radical selection of the proposed character (if any).

b. Alternative stroke count: Same reason, it would be nice to have alternative stroke count (if any)

c. Proposed location of the character within Super-CJK-TO-BE. This location data is named as “Pseudo Kang Xi index”

d. Data on the similar shape or same origin character(s) such that the review of “unify with already existing character or not” decision would be much easier. Also, the explanation why it should be independent (not unified) to be added.

e. If the characters are already used in the country by mean of local unique method, It is necessary to specify the method. This is just in case information to avoid possible interoperability problems.

f. In the ad-hoc discussion, needs of (verification) “tool(s)” are expressed. It would be nice if the proposals are including a data for the tools. (hopefully, in common format)

.

As a conclusion “ext.-C” ad-hoc group agreed to propose following data set as a necessary data for “ext-C” proposal submission.

a. Country index for extension-C proposal

b. KX radical with alternate radical (if any)

c. KX number of stroke with alternate number of stroke (if any)

d. Pseudo Kang Xi index

e. Source information

f. Glyph & Font Glyph shape in proposal should be large enough for review. 6/8 mm square is not enough (Sato’s personal addition)

g. Cognate with or Similar to information (with justification comment)

h. Temporary solution (if any)

i. Comparison data for tool (IDS?)

4. Tool development Joint team

The “ext.-C” group recognizes the usefulness of the “comparison/verification tool” for the ext.-C development. The “ext-C” group recommends to open a discussion between people who already have a prototype (or using) such a tool.

There are 5 tools (or something developing aid) are listed, “ext-C” group decided to hold ad-hoc discussion between owners of those tools. The owners are:

Macau Mr. Chau, C.K Clement, Chan, K.K. Arnnie

H.K. Dr. Lu, Qin

ROK Mr. Lee Keon Sik

TCA Pan, Cheng-Wu

China 2 tools TBD

The goal of discussion is to find out a possibility of “common data format” which to be supplied with the “ext-C” proposal by all IRG members for the tools.

There are explanations about the existing tools (including prototype and idea phase one), most of them are character search engine to find out if it is already coded or not (even if they are different each other). One image based idea is expressed by ROK, but it is almost in investigation phase. One common comment was a need of data for both evaluation and real use. Most of the data needed are already in SCJK. There are two kinds of tools. One is for “check if it is already coded”, and another is “compare if those are to be unified”. All tools expressed were the tool to check existing code table for preparation of the extension-C proposal.

The discussion concludes as:.

a. Machine readable data of Super CJK to be available for IRG members
b. If there is a search engine being used by IRG editor, it would be opened for IRG members for internal IRG members usage
c. IRG should encourage an free exchange of the available tools between the IRG member.
d. Each IRG members are encouraged to exchange an experiences and upgrade results of the tools
e. The IRG should encourage a e-mail discussion between the interested peoples on the tools
f. One of the key data is a component data. The ad-hoc group suggest to the IRG to consider IDC/IDS as a standard methodology to exchange the component data.
g. The group recognized that there is an image based tool ideafrom ROK, IRG may review the

tool for IRG use when it is available for evaluation at IRG.

5. Font issue

The objective to select new font format for a country submission is

“For ext-B, IRG had many review cycles due to the font data conversion. Reduce number of review cycle by keeping higher quality is necessary for timely release and better quality of printed code page of the ext-C”.

128 X 128 TT format is selected as a goal. 128 X 128 BM is interim if it is necessary.

6. Conclusion

6-1. Total proposal is not stable enough for IRG to review. It is better to wait for #18 meeting.

6-2. There will be a significant amount of proposed characters which are cognate with already coded characters. Special attentions, such as additional information within the proposal, review process…..might be necessary prior to the character review.

6-3. Still non-cognates are also in proposal. Therefore, review process would be not only one, but may be multiple process according to the characteristics of the proposed characters.

6-4. It is highly recommended for the countries to consider a re-submittal of a proposal with newly defined data-set.

6-5. IRG chief editor is recommended to consider making machine usable data of S-CJK and search engine available to the IRG members for ext-C development purpose.

6-6. It would be better to use higher quality font from the early stage of the extension-C review.

6-7. It would be better to open small gate for additional request from the country that already submitted the proposal. (No clear cut-off date for a while)

---end---2001-06-21 TKSato-Japan