


文字コード標準化における検討状況(文字数)について

2010/12/14

日立 山本

第 1 回委員会において、この事業で作成対象になっているフォントの文字数についての紹介があった。これを補足する意味で、標準化あるいは審議中案件の状況について報告する。



2.作成する文字（詳細）

種類	文字数
① ISO/IEC 10646 (UCS) 「BMP」	約2万4千文字
② ISO/IEC 10646 (UCS) 「Plane2 (Extension-B)」	約2万5千文字
③ ISO/IEC JTC1/SC2 へのISO/IEC 10646 (UCS) 「Extension C、Extension D、Extension E」提案により符号位置が定められているもの（現在、標準化手続きが進行中であるものを含む）	約1千文字
④ Unicode Consortiumへの「IVD」登録手続き中であるもの。	約2千文字
⑤ 上記①、②、③、④以外の漢字 (UCSの符号位置を持たず、ISO/IEC JTC1/SC2/WG2への符号化提案や Unicode Consortiumへの IVD登録手続きもされていないもの)	約7千文字
⑥ 国内規格の非漢字 「JIS X 0221:2007に含まれ、公的文書で一般に使われる非漢字」を以下のものと想定する。 (A)JIS X 0213収録の非漢字：1158文字 (B)JIS X 0212収録の非漢字：266文字（(A)との重複を含む）	約1千文字
⑦ 変体仮名 変体仮名については、汎用電子情報交換環境整備プログラムや情報処理学会情報規格調査会SC2専門委員会において、符号化標準の議論がなされておらず、また、参照すべき字体について公開された資料が存在しない。参照すべき字体、同定基準については推進委員会での検討結果に従うものとし、本提案における作業見積には含めていない。	168文字と想定されるが、後日調整する

10

① ISO/IEC 10646 (UCS) 「BMP」

この中には以下のものが含まれる

CJK Unified Ideographs：20,951

CJK Unified Ideographs Extension A：6,584

計 27,535

なお、BMPにはCJK Compatibility Ideographsに該当するものが457ある。

② ISO/IEC 10646 (UCS) 「Plane 2 (Extension B)」

CJK Unified Ideographs Extension B：42,711

なお、Plane 2にはCJK Compatibility Ideographs Supplementに該当するものが542ある。

③ ISO/IEC JTC1/SC2 への ISO/IEC 10646(UCS) 「Extension C、Extension D、Extension E」提案により符号位置が定められているもの（現在、標準化手続きが進行中であるものを含む）

CJK Unified Ideographs Extension C は追補 5 で収録された： 4,149

CJK Unified Ideographs Extension D は追補 8 に収録される予定だが、追補 8 はまだ発行されていない。ただし、Unicode 6.0 には既に収録されている： 222

CJK Unified Ideographs Extension E は現在 IRG での審議が中断している。審議の再開は早くても IRG#37(2011 年 11 月)以降になる見通し。現在の Extension E のドラフトには議論中のものも含めて約 6,700 字が含まれている

④ Unicode Consortium への「IVD」登録手続き中であるもの

汎用電子成果に基づく最初のバリエーション提案については 11 月 14 日付で正式に IVS の割当てが行われた。この中には 1,918 の基底文字(base character)に対して 4,195 のグリフをバリエーションとして登録した。

UCS	variant-1	variant-2	variant-3	variant-4	variant-5	variant-6	variant-7	variant
	(variant-11)	(variant-12)	(variant-13)	(variant-14)	(variant-15)			
03691	𠄎	𠄎						
	IA0204	JTAFBF						
0382F	𠄎	𠄎	𠄎					
	IA0254	KS099080	JD0879					
03836	帆	帆						
	JD0881	JTB09F						
03917	𠄎	𠄎						

⑤ 上記①、②、③、④以外の漢字（UCS の符号位置を持たず、ISO/IEC JTC1/SC2/WG2 への符号化提案や Unicode Consortium への登録手続きもされていないもの）

汎用電子委員会からこの事業に提供された文字一覧表において、登記固有文字にしか見られないものを除いた文字数(グリフの数)は 58,712。この中で

a) 属性情報として UCS の対応がつけられているもの： 55,933

b) a)のうち、UCS の値が重複するものを除くと： 50,246

c) a)-b)の 5,687 についてはバリエーションとして IVD に登録が必要な可能性がある。

d) UCS の対応がつけられていないもの： 2,779

c), d)について今後 10646 の統合漢字への追加、または Unicode IVD への登録申請が必要なものが含まれるとして、現在汎用電子委員会で精査を進めている。また、a)についても対応付けの妥当性について確認が必要との認識であり、その作業を通して c)や d)に該当するものが見つかる可能性がある。

⑥ 国内規格の非漢字

「JIS X0221:に含まれ、公的文書で一般に使われる非漢字」を以下のものと想定する。

(A) JIS X0213 収録の非漢字：1158 文字

(※X0213 の非漢字は 1183 字であるが、そのうち 25 字は結合文字である)

(B) JIS X0212 収録の非漢字：266 文字 ((A)との重複を含む)

⑦ 変体仮名

変体仮名については(2 文字の例外を除いて)文字コード標準には収録されていない。従って、何文字必要かについては不明。

以上