Vol.2011-CH-89 No.4
2011/1/22

Web コラボレーションサービスを利用した大規模漢字集合フォントの制作

上地 宏一†

Web 上で漢字字形を共有するデータベースを用いて Unicode(ISO/IEC10646)に対応するような大規模漢字集合フォントの制作が主にボランティアの手によって進行している。その進捗状況や制作過程における問題点と将来の可能性について述べる。

Producing a Large Kanji Characters Set Font Using Web Collaboration Service

Koichi Kamichi†

The production of the large Kanji character set font that corresponds to Unicode (ISO/IEC 10646) by using the database that shares Kanji glyphs on the web progresses by the volunteer.
This paper shows the progress report, problem and the possibility in the future.

† 大東文化大学
Daito Bunka University

1. はじめに

漢字字形を自由に登録・管理できるインターネット上に構築されたデータベース「グリフウィキ[1]」は公開を開始してから 3 年が経過し、順調な運用となっている。その途中経過についてはすでに報告している[2]が、その後の状況と大規模漢字フォントの制作について本稿では述べることとする。

2. グリフウィキの運用状況

2.1 レコード数

グリフウィキにはグリフ（漢字字形）のほか文章を登録することが可能で、両者を合わせたレコード数は 2010 年 12 月時点で 53 万レコードを超えている。グリフウィキは各レコードのデータ更新の結果、新旧すべての版を保存し、呼び出すことが可能であるため、この 53 万レコードという数字は実際には 16 万強のグリフページと 2000 余の文書ページから成り立っている。グリフページ数とその増加の推移を表したのが図 1 である。グリフの増加の度合いはまだ一定とは言えず時期により斑が見受けられるがおおむね順調に増加していると考えられる。

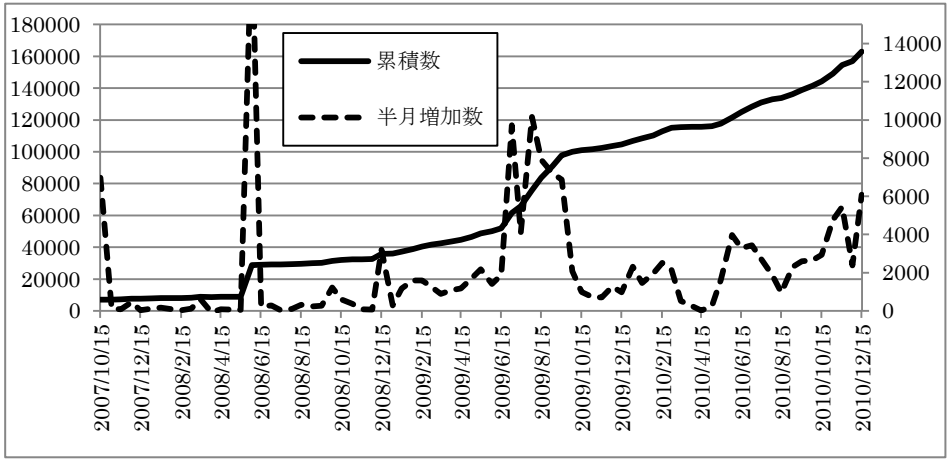

図 1 グリフページ数とその増加の推移

2.2 ユーザー数

グリフウィキは登録ユーザーと匿名ユーザーの 2 種類があり、登録の有無により機能に制限があるわけではない。これは積極的なデータベースの利用を促すための方針であり、他の同様のサービスのように IP アドレス（またはそのハッシュ値のようなもの）で同一投稿者の特定ができるようなこともなくサービス上は完全に匿名にて活動

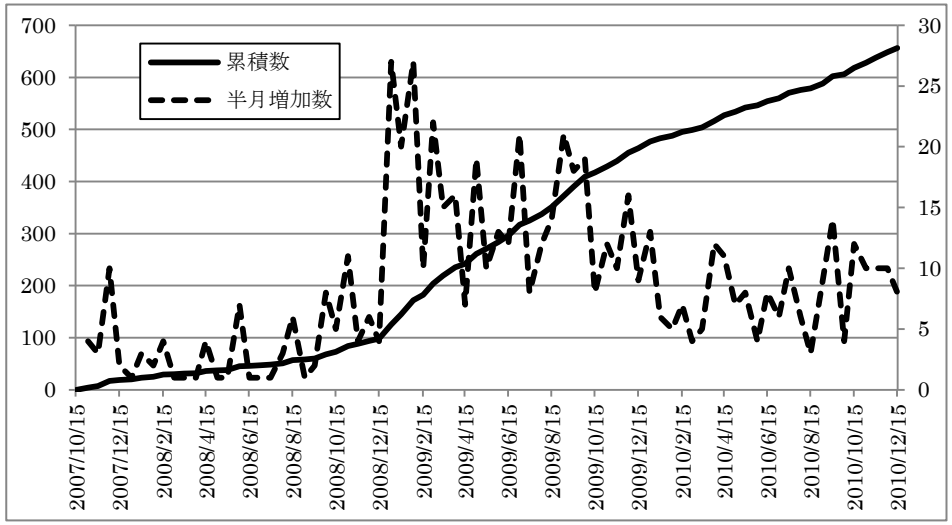

図 2 ユーザー新規参加数および累積数の推移

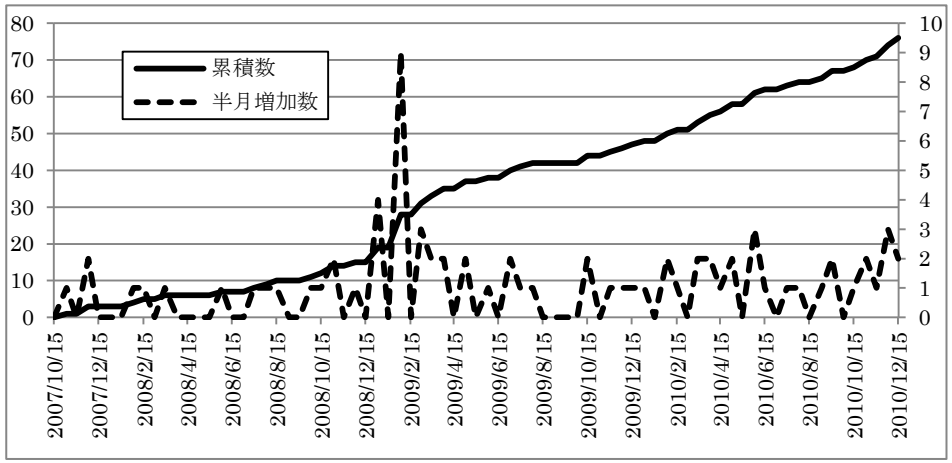

図 3 登録ユーザー数およびその増加の推移

できるようになっている。

ユーザー新規参加数および累積数の推移が図 2 である。ここで述べるユーザーとは匿名ユーザーも含み、アクセス元 IP アドレスをもとに同一かどうかを判断しているため、厳密なものではない（アクセスのたびに IP アドレスが変化するような環境では同一人物によるアクセスが複数人にカウントされる。また企業や大学などの機関内からアクセスした場合にプロキシーを経由しているとすべて同一人物としてカウントされる）が、2008 年末にピークに達して以降なだらかな減少傾向が見受けられる。

登録ユーザー数およびその増加の推移が図 3 である。現在 70 名を超える程度であり、決して多いとは言えない。増分も 1 か月に 2 名程度であるが、安定して増加していることだけは言えよう。

登録ユーザーのうち上位 21 名ごとの投稿数を表したのが図 4 である。この数は一括登録のような半自動的な投稿や、エイリアス（別名グリフ）の投稿はカウントせず、純粋なグリフデータの投稿数である。また 22 位以下の全登録ユーザーの合計投稿数、研究費による投稿数、および匿名ユーザーの投稿数も同時に列挙した。残念ながら上位 11 名により全投稿グリフの 75%を占めることになる。グリフウィキの利用は一般に開かれているが、実際にはある特定のユーザーの活動によってグリフ数の継続的な

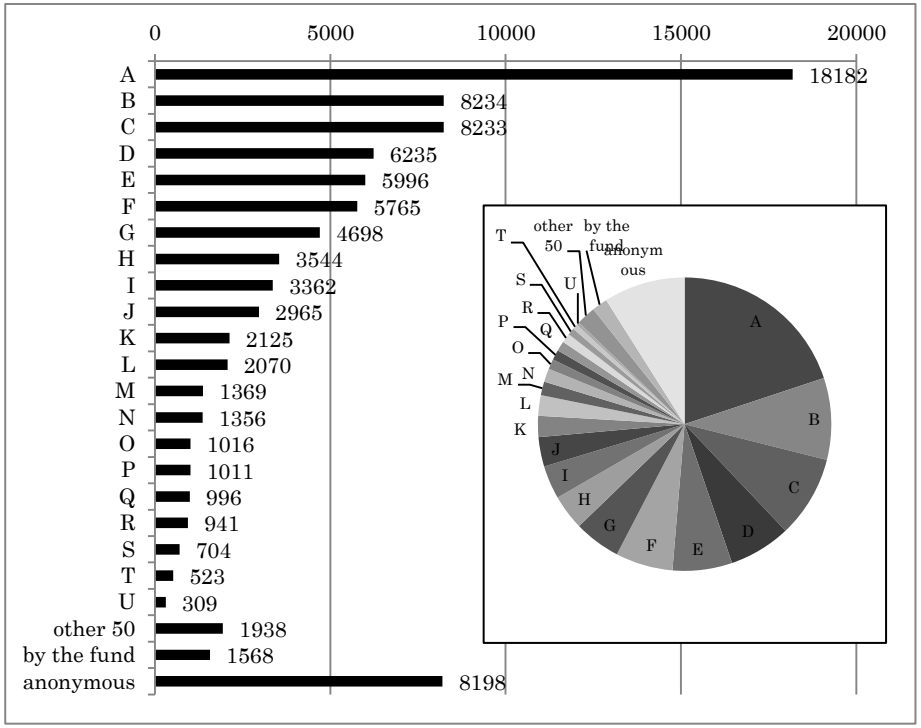

図 4 登録ユーザーのうち上位 21 名ごとの投稿数

Vol.2011-CH-89 No.4
2011/1/22

登録作業が行われている。ただし以前は上位 6 名で 75%を占めていたので、良い方向に分散しているものと考えられる。

また、同じく登録ユーザー上位 21 名の半月ごとの投稿グリフ数の推移を示したのが表 2 である。半月に投稿数が 100 を超えた場合、および 1000 を超えた場合にマス目を塗りつぶしている。この図から分かることは多くのユーザーが、一定期間に集中して大量のグリフを投稿しているということであり、逆に定期的に長期にわたって利用しているユーザーが少ないことである。あるユーザーは「グリフウィキには現実逃避的な中毒性がある」と述べていた。このことがユーザーの投稿活動の時期的特徴をよく表していると言える。また、グリフウィキの登録ユーザーは比較的時間の取れる学生・生徒が多い（あくまで本人の言及からの推定である）ということも特徴である。

3. 大規模漢字集合フォントの制作

筆者らは、グリフウィキに登録されているグリフを利用してフリー（無償・自由）の大規模漢字フォント（「花園明朝[3]」）を公開してきた。グリフウィキの初期データこそ研究助成を利用して作成したデータであるが、その後はボランティアのユーザーの投稿によりグリフ数を増やしてきた（今年度より科研費を利用したデータ作成も開始した）。大規模漢字集合というのは具体的には ISO/IEC 10646[4](Unicode、以下 UCS と記す)のことであり、すでに 7 万余字の漢字集合となっている。「花園明朝」の収録文字数の推移は表 1 の通りである。途中までは国内 JIS 規格の漢字集合に限っていたが、2009 年 9 月の版より収録対象を UCS 集合に広げたため一気に字数が増加した。2009 年 12 月からは IVD[5]集合も収録対象に加えている。

グリフ数	公開年月
6,356	2007 年 6 月
10,204	2008 年 10 月
14,368	2009 年 3 月
17,825	2009 年 5 月
48,816	2009 年 9 月
52,809	2009 年 12 月
57,557	2010 年 7 月
60,000	2010 年 10 月

表 1 「花園明朝」の収録文字数の推移

本来グリフウィキは、ユーザーが自分で必要なグリフを登録し公開するものであり、UCS 集合が拡充されることは直接の目標ではない。しかしながら UCS 集合が充実することにより、そのグリフを加工したり、部品を組み合わせたりすることでより平易に新しいグリフが作成可能となる。また、定期的に「花園フォント」として公開する

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
2007/10/15																					
2007/11/1																					
2007/11/15																					
2007/12/1																13					2
2007/12/15																37					
2008/1/1																32					
2008/1/15																					
2008/2/1																17					
2008/2/15																					
2008/3/1																58					
2008/3/15																54					
2008/4/1										15						34					
2008/4/15										24						8					
2008/5/1										50						7					
2008/5/15										42						5					
2008/6/1										279						12					
2008/6/15										121						18					
2008/7/1										48						23					
2008/7/15										8						20					
2008/8/1		30								69						7					
2008/8/15		153								123						50					
2008/9/1		11								177						58					
2008/9/15		123								94						71					
2008/10/1		443								19			794			165					
2008/10/15		283								18			575			112					4
2008/11/1		77								53						53					
2008/11/15		25								13						15					
2008/12/1		42								11											
2008/12/15		43								4											
2009/1/1	24	178								30						1				25	
2009/1/15	316	26								4						33				29	
2009/2/1	436	625	424				14			3				118		15				103	33
2009/2/15	355	437	693							19				166		5				136	85
2009/3/1	332	262	525			89				3				95	6	17				136	79
2009/3/15	173		460			65				19				106	52	4		13			31
2009/4/1	309	332	184			142				29				179		1	44	135			10
2009/4/15	211	658	347			283				18				45		9	82				27
2009/5/1	595	971	445			547				9				57			618				5
2009/5/15	772	661	292			721				104				156	1		252				11
2009/6/1	306	245	28			380				55				81	11	9					
2009/6/15	252	443	95			380				78				20	16	12		2			
2009/7/1	151	184	28			285				198				33	6				1		
2009/7/15	276	752	275			336				101					14			69	320		22
2009/8/1	249	475	194	106		681				34				108	2	3		306	316		
2009/8/15	235	516	46	595		289				14				23	7	1		31	56		
2009/9/1	596	239	274	351		407				34				52		5		77	8		
2009/9/15	434		20	107		709	1			38				18	8	3		24			
2009/10/1	681		320			378				43					9			16	1		
2009/10/15	293		24	34		54				14				49	12			99			
2009/11/1	209		7	185		19				33					12			46			
2009/11/15	278		456	254						38				8		1		4			
2009/12/1	338		753	500						152				6							
2009/12/15	248		230	78				110		120				2	44			42			
2010/1/1	846		199	60				325		151					65	8				2	
2010/1/15	565		34	201				248		82					72						
2010/2/1	932		13	253				385		23									2		
2010/2/15	1117		256	1104				816	1	16					45	5					
2010/3/1	4467		161	1017				1138		7					4	7		33			
2010/3/15	2047		25	184				411		4					72			13			
2010/4/1	139			48				107		7					95			20			
2010/4/15				12			1	1		2				11				1			
2010/5/1				92			11	3						23	4			3			
2010/5/15			69	1			9								1						
2010/6/1			23	151			733		31	3		5						1			
2010/6/15				175			234		135			79			84			3			
2010/7/1			58	134			455		194	9		404			87						
2010/7/15				15			26		139	14		506			9						
2010/8/1				85			20		206	29	119	475			115	2		3			
2010/8/15				14			4		157	7	9	140			13						
2010/9/1				2			235		328	16	41	21			9						
2010/9/15			20	33			369		339	111	27	100			20						
2010/10/1			389	25			473		365	32	439	83			1					8	
2010/10/15			26	98			348		510	24	643	83									
2010/11/1			271	98	2425		293		265	59	177	135			45					49	
2010/11/15			470	107	3060		140		208	7	267	39			37					35	
2010/12/1			64	68	433		98		137	3	82				1	1					
2010/12/15			35	48	78		1070		327	1	320				35						

表 2 登録ユーザー上位 21 名の半月ごとの投稿グリフ数の推移

ことにより、グリフウィキの活動を広くアピールする手段になっている。

「花園フォント」は筆者が運用するウェブサイトで公開しているだけでなく、Linux の複数ディストリビューションにおいてパッケージとして登録されている。

3.1 漢字集合の充足度の推移

図 5 は UCS に含まれる部分集合のグリフウィキ（および「花園フォント」）での充足率の推移である。それぞれの集合は数量が異なるため一概に比較はできない。すでに Ext.B（CJK 統合漢字拡張 B 集合）を除くすべての集合を網羅している。最後に残っている Ext.B 集合は現在約 74%を充足しており、残りは 11,000 余字である。このほかに、現在審議中の Ext.E 集合や先般追加された IVD 集合が将来的には追加されることとなる。

現在 Ext.C や Ext.D を収録したフォントは非常に少なく、グリフウィキの活動および「花園明朝」の公開は文字コードの標準化活動にとっても非常に意味のあることだと確信している。

3.2 問題点と検討事項

3.2.1 デザインの統一性

デザインの美的センスは主観によるところが大きく、またデザインポリシーによっても左右される。フォントデザインに当てはめると、たとえばフトコロの広さや重心の取り方はデザイナー・フォントによって変わる。一方グリフウィキはだれでもグリフを登録することが可能である。このため、ユーザーによって異なるデザイン感覚でグリフ投稿がなされると全体として統一性のないフォントになってしまう可能性がある。

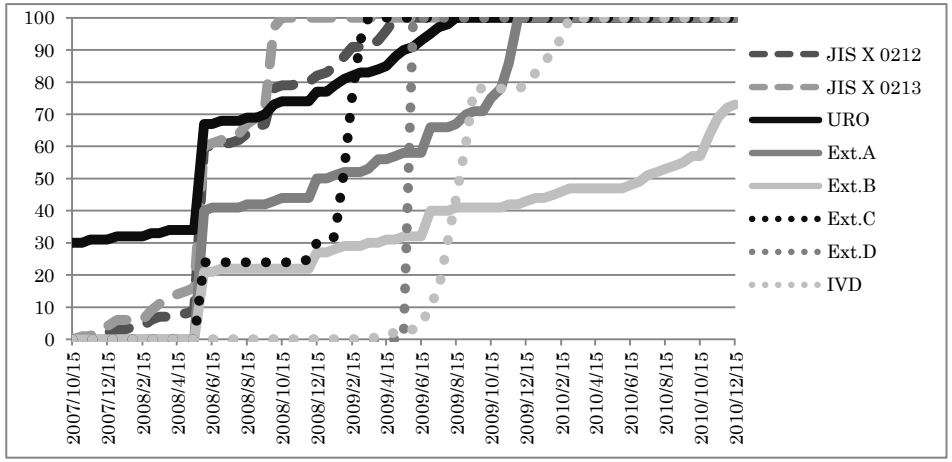

図 5 UCS に含まれる部分集合のグリフウィキでの充足率の推移

しかしグリフウィキでは、積極的に部品の活用がなされ、なるべく既存の部品を組み合わせて新しいグリフがデザインされている。グリフウィキでは「人偏」「走ニョウ」などの部品を「偏化変形」グリフと称しているが、これらグリフが 1600 程度登録されている。これら部品を利用して文字をデザインすると結果的に部品のデザインに影響され結果的にデザインの統一性が保たれていると考えられる。

3.2.2 UCS の抽象性

UCS の漢字集合部分は、複数の地域規格等の組み合わせであり、1 つのコードポイントに複数の異なる字形が Unify されている。たとえば「骨」は中国大陸の字形と日本の字形が正反対となる有名な例である。グリフウィキはこの UCS のコードポイントをそのままページ名として利用している。前述の「骨」の場合、コードポイントは U+9AA8 をベースとした「u9aa8」がグリフウィキでのページである。この「u9aa8」ページに登録されているグリフは日本の規格票に沿った字形が登録されている。一方中国大陸の字体は「u9aa8-g」というページに登録され（図 6）、ほかに「-t（香港・台湾）」、「-k（韓国）」、「-v（ベトナム）」などの接尾語が用意されている。

図 6 U+9AA8 に相当するグリフ（左：u9aa8、中：u9aa8-g、右：u9aa8-t）

グリフウィキは UCS のコードポイントをそのままページ名としたグリフ（以降無印 UCS グリフと呼ぶ）に対しては、これまで以下の 2 種類の字形ルールを設けてきた。

- 第 1 ルール：日本国内規格で規定されている場合はその規格票の字形に沿う
- 第 2 ルール：日本国内規格で規定されていない場合はなるべく仮想 J 字形とする

仮想 J 字形というのは、もしそのグリフが JIS 規格に収録される場合に想定される字形のことで「平成明朝体‡」を想定している。このような方針とする理由は、フォントとしてまとめた場合に文字によって字形が一定しないのはおかしいと考えるからである。たとえば文字によって草冠が 3 画であったり 4 画であったり、などの不統一を排除することを目的としている。結果的にグリフウィキは日本デザインを最優先する、ということになる。

しかしこの方針は日本人にはおおむね受けいれられるルールであるが、外国のユーザーにとっては受け入れがたい場合もある。彼らの言い分は「漢字は日本人のものだけではない」であり、もっともである。グリフウィキは日本語だけでなく英語のユー

‡ 当初、財団法人日本規格協会文字フォント開発普及センターで開発されたフォント

ザーインターフェースも用意されている。このほか将来的には中国語や韓国語も提供予定である。外国人のユーザー（あくまで本人の言及からの推定）も現に存在する。

先般、登録されている無印 UCS グリフに対して、日本デザインから他国・地域デザインへの改変がなされたり、その逆の改変がなされたりするようになった。広い意味での編集合戦が生じつつある。

そこで大きく方針を変更し、無印 UCS グリフを廃止することを検討している。つまり無印 UCS グリフをすべて「-j」と「-jv」に移行させる。日本国内の規格票にあるものは「-j」とし、無いものは従来通り「仮想 J 字形」として「-jv」を割り当てる。無印 UCS は消滅するので日本人は「-j」、「-jv」に登録し、他の国・地域の字形を登録する場合は「-g」、「-t」、「-k」、「-v」等に登録することで平等が図られる。

また、「汎用電子情報交換環境整備プログラム§」で規定された 7 万弱字の漢字集合は、その多くが UCS と対応付けることが可能である。そしてその字形はすべて平成明朝体で実装されている。そこで「-jv」字形はこのプログラムで対応付けられた字形を根拠とすることが望ましいと言える。

3.2.3 フォントファイルの問題点

現在のコンピュータ用フォントは TrueType 形式や OpenType 形式が利用されている。これらは 1 つのファイルに収録できるグリフ数の制限が 65000 余字(16bit)となっているため、すべての UCS グリフを収録するためにはフォントファイルを 2 つ以上に分割する必要がある。拙稿[6]で述べたようにユーザーのニーズによって 2 つのフォントに分化することを検討している。

4. おわりに

現在のグリフウィキの投稿ペースが維持されるならば UCS の Ext.B 集合はあと 1 年程度で完全収録が達成できると予想している。ボランティアベースの作業によってこれほど大規模の漢字集合フォントが制作されたことは過去にない。ウィキという新しい手段によって大きな成果が得られる 1 つのモデルケースになることに期待したい。

付記　本発表は科研費（課題番号 22700262、代表者：上地宏一）による成果の一部を含むものである。

参考文献

1) 上地宏一, 「漢字グリフ管理 Wiki システム（GlyphWiki）の構築」, 『人文科学とコンピュータシンポジウム論文集』(じんもんこん 2007), pp.237-244, 2007.
2) 上地宏一, 「漢字字形管理環境 GlyphWiki(グリフウィキ)」, 『東洋学へのコンピュータ利用 第 19 回セミナー』, pp.127-141, 2008.
3) 上地宏一, 師茂樹, 「自由な漢字フォント環境の構築に向けて」, 『東洋学へのコンピュータ利用 第 17 回研究セミナー』, pp.121-127, 2006.
4) ISO/IEC 10646:2003 Information technology -- Universal Multiple-Octet Coded Character Set (UCS)
5) Unicode Technical Standard #37 - Ideographic Variation Database
6) 上地宏一, 「フォント・ブラウザ・多漢字」,ソフトウェア・レビュー, 漢字文献情報処理研究第 11 号, pp.127-137, 漢字文献情報処理研究会, 好文出版, 2010.

§ 経済産業省の委託事業として情報規格調査会，独立行政法人国語研究所，財団法人日本規格協会の三者により平成 14 年度から 4 年間実施されたプロジェクト