論文

国際標準化文字集合に対応した大規模漢字フォントの制作

上地　宏一

2023年9月に公開された国際標準化文字コードUnicodeに収録される10万弱の漢字およびIVDに登録されている3万弱の異体字を全て収録する大規模漢字フォントを制作した。加えてその13万字種を検索できるツールの制作を行った。本稿ではその詳細を述べる。

キーワード：Unicode、漢字、異体字、フォント、漢字部品

１．はじめに

現在、コンピュータで文字を扱うための基盤となる文字コードはUnicode標準（および国際標準規格であるISO/IEC 10646。以降まとめてUnicodeと記す）にほぼ統一されたと言っても過言ではないが、Unicodeはこの30年間に繰り返し更新され、字種が追加されてきた。本稿では、Unicode 15.1版（2023年9月公開）に対応した大規模漢字フォントの制作とその公開について述べる。

２．背景

2-1　Unicodeの漢字集合の現状

Unicodeに収録されているCJKV統合漢字・互換漢字およびIVDをまとめたものが表1である。Unicodeの改版に伴い、漢字ブロックが追加され、収録される漢字集合が増えている。2024年1月現在最新となる15.1版では98,682漢字が収録され（Unicode 2023a）、またIVDにおいて29,437グリフが登録されている（Unicode 2022）。例えば日本で用いられる漢字は常用漢字の2,136字や人名用漢字の863字などがあり、一般的な漢和辞典の親字数は1万字強である[1)]のに対し、Unicodeには大型漢字字典に収録される史料的な漢字や異体字が多く含まれ、また中国大陸、台湾、韓国、ベトナムなどからも継続的に字種追加の申請がなされた結果、ある種の博物館的文字コードになっている。Unicodeの収録文字に対する漢字の比率は高く、15.1版の全149,813文字に対して、純粋な漢字だけでもその65%を占めることになる。

表1　Unicodeの漢字ブロック一覧（Unicode 2023bをもとに筆者作成）[2)]

ブロック名	コードポイント範囲	文字数	Unicodeの版番号と年
URO	U+4E00 ～ U+9FFF	20,992	1.1版 1993年
拡張A集合	U+3400 ～ U+4DBF	6,592	3.0版 1998年
拡張B集合	U+20000 ～ U+2A6DF	42,720	3.1版 2000年
拡張C集合	U+2A700 ～ U+2B739	4,154	5.2版 2008年
拡張D集合	U+2B740 ～ U+2B81D	222	6.0版 2009年
拡張E集合	U+2B820 ～ U+2CEA1	5,762	8.0版 2012年
拡張F集合	U+2CEB0 ～ U+2EBE0	7,473	10.0版 2015年
拡張G集合	U+30000 ～ U+3134A	4,939	13.0版 2019年
拡張H集合	U+31350 ～ U+323AF	4,192	15.0版 2021年
拡張I集合	U+2EBF0 ～ U+2EE5D	622	15.1版 2023年
互換漢字	U+F900 ～ U+FAD9	472	――
互換漢字補助	U+2F800 ～ U+2FA1D	542	――
IVD	――	29,437	2022-09-13版

2-2　大規模漢字フォント

Unicodeの拡張によってさまざまな漢字の情報交換が保証されるようになっても、実際に画面に表示したり、紙に印字したりするためにはフォントが必要になる。漢字のような数の大きな文字集合ではフォントの制作にコストがかかるため、欧文と比べてそのバリエーションは少なくなる。さらにすべての漢字集合を収録する必要はないため、例えば日本の漢字フォントには中国の簡体字が含まれなかったり、あるは史料的な漢字が含まれなかったりするなど、文字コードに収録されていても実質的に「使えない」漢字というものが存在し、あるいは「使えない」漢字の方が多いのが現状である。

その中で「大規模漢字フォント」と呼ばれるものが存在し、Unicodeの10万の漢字を部分的に網羅することで、珍しい漢字をコンピュータで処理したいという一部のニーズを満たすことが可能となっている。そのいくつかを列挙したのが表2である。これらのほとんどはOSに標準で添付されていないため、使いたい人は自分でフォントをダウンロードしてインストールする作業が必要となる。

表2　有名な大規模漢字フォント

名称	説明
Simsun-ExtB・MingLiU-ExtB	Windows Vistaから標準搭載された拡張B集合対応フォント
今昔文字鏡・GT 明朝・T フォント	それぞれ数万字を収録。Unicode に準拠していないため情報交換には不適
IPAmj明朝[3)]	日本の行政処理で収集・整理された文字を収録した6万字のフォント。Unicodeのコードポイントに対応していない漢字は未収録
BabelStoneHan[4)]	最新の拡張I集合までを網羅（全てではなく一部）する6万字を収録。Unicode収録漢字のカバー率は57%

花園明朝[5]	筆者により公開。拡張F集合まで収録、2017年以降更新停止
謎乃明朝[6]	グリフウィキの成果を利用し、花園明朝の後継ともいえる。拡張I集合を含むUnicodeの全漢字を収録するが、非漢字は未収録

2-3　花園明朝フォントとグリフウィキ

「花園明朝」フォントは筆者によって公開されている大規模漢字フォントである。Unicode の漢字を中心に収録したフォントであり、2007年6月に公開を開始して以来、収録文字数を増やしながら版を重ね、2017年の最新公開版では拡張F集合までの10万字（非漢字含む）を収録した。世界でも数少ないUnicodeの大規模漢字集合を収録したフリーフォントとなっている。この花園明朝フォントの制作には「グリフウィキ」と名付けられた漢字字形共有Webデータベースのグリフデータが活用されている。グリフウィキも筆者が運営しているWebサイトであり、2007年10月の公開以降、多くのボランティアによって100万を超えるグリフが登録されている。Unicode の漢字集合の追加はISO/IEC 10646のJTC1/SC2/WG2委員会の中のIRGで審議されるが、その議論の過程で扱われる文書・資料はWebで公開される[7]。この資料を基に追加の候補となる漢字をボランティアが同時進行的にグリフウィキに登録することにより、文字コードの新版の公開の時点で（あるいは公開後、そう遅くない時期に）、グリフウィキにその追加字種データが整備されていることが実現している。

筆者は定期的にUnicodeの漢字ブロックの拡張に対応して花園明朝フォントの版の改訂を行ってきたが、2017年9月の改訂以降、いくつかの問題により更新が止まってしまった。一番の原因は、2009年からグリフウィキに登録されている非漢字を積極的に収録することにした方針変更である。もともとグリフウィキ（および内部で使われている漢字字形生成エンジンであるKAGEシステム）は漢字字形を扱うものであったが、必然的に非漢字も登録されてきた。公開当初の花園明朝フォントは「漢字」のフリーフォントを謳っていたが、文字種数の多さをアピールする狙いもあり非漢字の収録を行うことになった。しかし、非漢字集合は多種多様のブロックで構成されていることや、各ブロックがどれだけの字種をカバーすれば、そのブロックを収録しているとアピールできるのかの知識が乏しく（例えば「ひらがな」ブロックを収録したとしても「「ぬ」だけは入っていません」では使い物にならない）、これらの管理を行うことが難しくなった結果、更新が滞ってしまった。

3．字雲（Jigmo）フォントの制作

3-1　概要

そこで筆者は、非漢字の収録をあきらめて漢字に特化した新たな大規模漢字フォント「字雲（Jigmo）フォント」を制作することにした。ただし収録対象を厳密に漢字だけに限定すると使いにくく、最低限の英数字やかな文字は必要であるため、収録する文字種を東アジア文化圏の文字に限定するとともに、フォントだけでなくフォントを生成するツールを併せて公開することで、利用者が自分のニーズに合わせて文字種の増減を行い、自身でフォントを生成できるようにすることとした。

3-2　フォント生成ツール

字雲フォントはフォント生成ツールとフォントファイルから成り立っている。ツールは以下の手順

でフォントを生成する。

1. unicode.org から文字コードのメタデータを取得
2. 利用者が定義したキーワードリストに基づいて、フォントに収録する文字種を決定
3. 文字種に基づいてコードポイントリストを生成
4. リストに基づいて各コードポイントの SVG データをグリフウィキから取得
5. FontForge でフォントファイルの生成
6. グリフ一覧（HTML）の生成

2. のキーワードリストは標準では図１のようなテキストファイルになっていて、IDEOGRAPHIC（漢字ブロックに記述される）によって漢字を収録し、CJK（中国・日本・韓国に関する文字につく）、HIRAGANA（日本語のひらがな）、HANGUL（韓国語のハングル）や BOPOMOFO（中国語の注音記号）など、東アジア文化圏の文字種を取り込むように記述している。これは Unicode が公開しているメタデータ[8]を使って Unicode 収録文字の文字名に部分一致したものをフォントへの収録対象とする仕組みとなっている。たとえばここに「TIBETAN LETTER（チベット文字の文字ブロック名）」を追記すれば、U+0F00〜U+0FFF に収録されるチベット文字が追加される。また、ASCII の文字（英数字・基本的な記号）については強制的に収録することとしている。

```
CJK
IDEOGRAPHIC
KANGXI
ROUNDED SYMBOL
IDEOGRAPH
ERA NAME
VERTICAL KANA
HIRAGANA
KATAKANA
HENTAIGANA
<square> 30
HANGZHOU
BOPOMOFO
MAHJONG
XIANGQI
HANGUL
Hangul
KOREAN CHARACTER
```

図１　収録対象キーワードリスト（keywords.txt）

このツールはあくまでもフォント生成に必要なデータとスクリプト（フォント生成の指示プログラム）を生成するだけであり、実際にフォントを生成するためには外部アプリケーションである「FontForge[9]」を利用する。ツールはPython言語およびPerl言語からなる複数のスクリプトとそれらを統括するシェルスクリプトの集合体である。

3-3　フォントファイル

現在のフォントファイル形式（TrueType および OpenType 形式）は内部テーブルの制約により、1つのフォントファイルに含めることのできるグリフ数が65,536個に限定されているため、Unicodeの全10万漢字を1つのフォントファイルに収めることができない。現在、Unicodeの漢字は第0面、第2面、第3面に収録されている（一部「かな」や記号・絵文字の一部は第1面）が、すでに第2面は6万字が埋まっていて、今後の字種追加も見越した結果、以下のようにそれぞれの面ごとにフォントファイルを分けた3ファイル構成とすることにした。

・Jigmo.ttf　第0, 1面の文字を収録（非漢字：ASCII、かな、記号、漢字：URO、拡張A、互換漢字）
・Jigmo2.ttf　第2面の文字を収録（非漢字：ASCII、漢字：拡張B, C, D, E, F, I)、互換漢字補助）
・Jigmo3.ttf　第3面の文字を収録（非漢字：ASCII、漢字：拡張G, H）
（※それぞれ、その面の漢字を親字とするIVDグリフを含む）

この結果、利用時に文字によって3つのフォントを切り替える必要が出てくるため不便であるが、現状では解決不能なため、将来の文字処理システムの改善（たとえばいくつかのフォントをまとめた論理・抽象フォント）をOSに期待するしかない。

3-4　GitHubでの公開

フォント生成ツールはGitHubのプロジェクトとして公開し[10]（図2)、また生成したフォントファイル（Jigmoフォント）も同プロジェクトのWebページに公開している[11]（図3)。フォントファイルのライセンスは CC0[12]（日本の著作権法では無効となるが、便宜的なパブリックドメイン相当）として自由に使うことができる。フォント生成ツールのライセンスはMITライセンス[13]である。

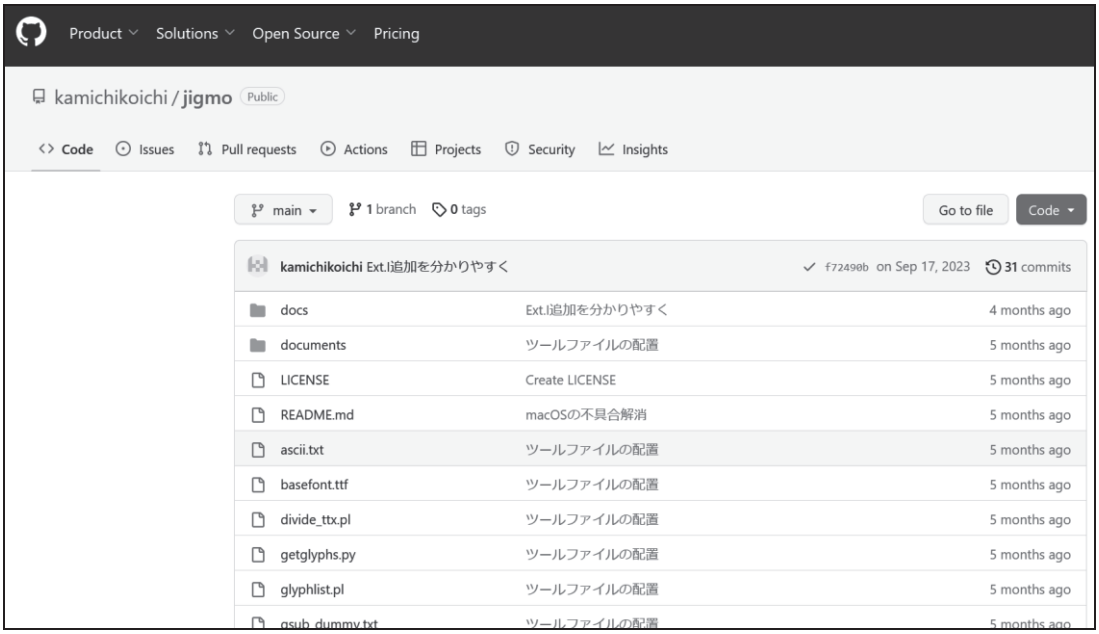

図2　GitHub でのフォント生成ツールの公開

図3　GitHub でのフォントファイルの公開

４．入力補助ツールの制作[14)]

4-1　多漢字の検索手段

字雲フォントの公開によってUnicodeのすべての収録漢字に対応したグリフ表示が可能となる一方で、その文字をどのように入力するのかという問題は解決していない。通常、我々がパソコンに漢字仮名交じり文を入力するときにはローマ字入力を用いる。これは漢字の「読み」を入力し、その読みに相当する漢字仮名交じり表記の候補を選択・確定するものである。しかしながらUnicodeの10万字種の文字を読みで入力することは効率的ではない、例えば「コウショウ」という読み仮名に対応する漢字熟語は48種類あるという（日本漢字能力検定協会 2021）。また古典文献に見られる漢字によっては、地名として記述されているだけで、読み（や字義）が不明なケースもある。従来、こういった大規模漢字集合から任意の文字を探し出す方法としては、以下の3系統に分かれるが、現実的な解として②の漢字部品検索が最適である。

①部首・画数から調べる

（・部首画数検字：部首と、部首を除いた画数（部首内画数）から文字を探す伝統的な検字方法である。214種類の部首は明代の漢字字典である『字彙』（明・梅膺祚編、1615年）、『正字通』（明・張自烈編、1627年）から清代の康熙字典（清・張玉書ほか編、1716年）に継承され、現代の漢和辞典・漢字字典でもこのいわゆる「康熙部首」ないしその派生部首が使われる。しかし同じ部首・同じ画数でも大量の漢字があり探すのは困難であるほか、Unicode 漢字特有のUnification（字形統合）により、同じコードポイントの文字であっても国・地域によって画数が変わることがあり、画数による検字を困難としている）

・規格票およびUniHanデータベースから調べる：規格票では漢字ブロックごとに康熙部首・画数順に並んでいるため、上記部首画数検字が可能であるが、漢字ブロックはURO、拡張A～Iの10パートに分かれているため、最悪の場合10回検字作業を繰り返すことになる。Unicodeでは10パートの漢字全てをまとめた上で部首画数順に並べたUniHanデータ[15)]およびその検索Webサービス[16)]を提供しているため、多少は労力を減らすことが可能となっている。

②漢字部品から調べる

・CHISE IDS Find[17)]：守岡知彦氏提供のWebサービスで漢字部品を入力し、その部品を含む漢字の一覧を検索可能。インターネットに接続している必要があること、部品を入力できる漢字スキルが必要である。

・文字情報基盤検索システム[18)]：一般社団法人文字情報技術促進協議会提供のWebサービスで漢字部品から文字を検索できるほか、既存の文字を分解して検索部品に用いるなどの工夫がみられる。インターネットに接続している必要があることと、経済産業省実施の文字情報基盤整備事業によって収集された日本の行政で使われた文字が検索対象のため、Unicodeの10万字すべてを検索できないことが残念である。

③手書きで調べる

・グリフウィキ手書き検索[19)]：kurgm氏提供のWebサービスで、グリフウィキに登録されている32万弱のグリフを手書きで検索できる。書き方に依存すること、Unicodeに関係ないグリフも検索

されること、インターネットに接続している必要があり、かつサービスがスリープ状態に入ると復帰に30秒以上待たされることなどがデメリットである。

4-2　IDSデータの活用

Unicodeの収録漢字は、IRGにおける追加収録審議の際にIDS（漢字構造記述）データを提出することになっていて、10万字すべての漢字についてIDSデータが確保されていることから、このIDSを活用した漢字部品検索ができないかと考え、入力補助（漢字検索）ツールを作成することにした。実際には拡張Ⅰ集合のIDSデータが探し出せなかったため、cwhk氏がグリフウィキに登録したIDSデータを含め、以下のIDSデータを組み合わせて利用することとした。

・CHISEプロジェクト公開のIDSデータ[20]：先述のCHISE IDS Findの元にもなっている、URO および拡張A～H集合（拡張Ⅰ集合が不足する）のIDSデータ。大元は台湾中央研究院のCDPデータベースをその一部としている。一部Unicode以外の漢字部品を用いているため、その部分は活用できない。

・漢字データベースプロジェクト公開のCJKVIデータベース[21]：CHISEプロジェクトのIDSデータをベースにしつつ、極力Unicodeの漢字のみで構成しているほか、CJKVの各代表字形にIDS上の差異がある場合は併記している（例：U+6F22「漢」：⿰氵⿱廿口夫[GTKV]　⿰氵英[J]　の2併記）。URO および拡張A～F集合（拡張G, H, Ⅰ集合が不足する）を網羅する。

・BabelStone公開のIDSデータ[22]：先述の「BabelStoneHan」フォントの公開者でもあるイギリスの中国学者であるAndrew West氏公開のIDSデータである。漢字データベースプロジェクト公開のCJKVIデータベースをベースにしつつ、IDSで表現できない3種の抽象表現記号（図4）を取り組んだ意欲的なデータである。URO および拡張A～H集合（拡張Ⅰ集合が不足する）を網羅する。

・cwhk氏がグリフウィキに登録したIDSデータ[23]：現状で唯一入手可能であった拡張Ⅰ集合のIDSデータである。

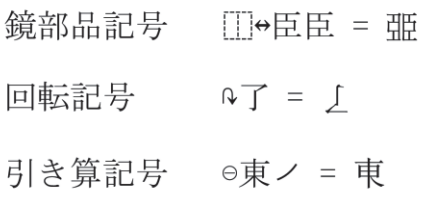

図4　3種の特殊記号

入力補助ツールは単一のHTMLファイルで構成されるスタンドアロン型のWebアプリケーションとし、ネットワークを必要とせずに内部に各種データを保持し、またJavaScript言語によってロジック部分を記述することする。これにより、HTMLファイル1つを取得すればツールが利用できる手軽さを実現できる。

4-3　漢字部品の定義

漢字部品を使った漢字検索には、その部品について2種類の考え方がある。

①任意の漢字部品を指定：あらゆる漢字部品を入力して検索するタイプで、先述の「CHISE IDS Find」や「文字情報基盤検索システム」が採用している。問題点は、漢字（部品）に対する知識の有無が検索に影響することである。

②部品セットの定義：漢字部品を100〜200種類用意し、その中から指定するタイプである。①と比較すると漢字に対する知識の必要性は低いが、部品数を絞ると検索結果の候補が多すぎることになり、逆に部品を多くすると部品の選択が難しくなる。

今回は②の方法を取ることとした。その理由はツールを1つのHTMLファイルで構築する目的があり、よりシンプルなロジックとするためである。部品の選別は以下の手順で行うこととした。

段階1　IDSデータを元に、全ての漢字部品とその出現回数をカウントする
段階2　設定した閾値で頻度の高い部品に絞り、それ以外の部品をIDSデータを使って分解する
段階3　再度、漢字部品とその出現回数をカウントし、閾値で頻度の低い部品を捨てる

閾値を106個とした場合の、段階1の時点でのカウント結果の上位20位を表3の左から1番目に、閾値前後10位を2番目に示す。網掛けは閾値以下で除かれる部品である。また、段階3の上位20位を表3の左から3番目に、閾値前後10位を4番目に示す。当初「鬼」「攴」「衣」といった部品は閾値より上位にあったが、分解後再整列の結果、閾値以下となっている。また再整列の結果、「一」が5万回を超える（10万字種の半分に含まれる部品であることを意味する）など、情報量としてやや問題があると言えなくもない。なお、この出現回数は同じ部品が2回以上出現する場合も1回とカウントする（「林」も「森」も出現する「木」部品は1回）。また抽出された106個の品を図5に示す。

また、検索に於いて「孫引き」を有効とするかどうかも検討の対象となった。例えば「口」という部品を含む漢字（例：「名」）を、「口」に含まれる「一」でも検索できるべきかどうかである。孫引きを有効とすると、特に画数の少ない漢字部品の情報量が低下するため、今回は無効とした。一方で、孫引きができないため、利用者はより高度な（複雑な）部品を指定しないと検索漏れが発生するという問題も生じる。今回は部品の個数を106個に抑え、画面上で一覧できることによりこの問題を回避できると考えた。

表3　部品の出現回数リスト

部品	出現回数	部品	出現回数	部品	出現回数	部品	出現回数
口	5,029	己	398	一	51,353	酉	896
⺾	4,175	厶	398	ノ	42,911	ネ	895
木	3,782	子	392	丨	35,075	水	894
氵	3,557	鬼	384	口	22,102	毛	799
金	2,752	攴	379	丶	17,308	羊	789
才	2,542	衣	377	乚	15,091	乃	783
土	2,515	刀	371	十	13,886	見	711
一	2,191	水	370	日	11,531	歹	683
火	2,161	羊	369	土	10,383	仌	683
日	2,144	風	367	人	9,710	革	674
糸	2,054	廾	365	八	9,440	衣	657
月	2,036	卜	365	冂	8,923	舟	601
亻	2,012	瓦	364	二	8,883	阝	590
山	1,966	勹	357	丷	8,801	走	589
女	1,922	骨	356	木	8,588	九	575
言	1,918	夕	355	冂	7,713	巳	566
竹	1,879	工	354	⺾	7,153	鬼	465
虫	1,840	黒	353	大	6,592	髟	460
鳥	1,813	儿	352	艹	5,984	入	458
魚	1,595	干	342	月	5,083	攴	450

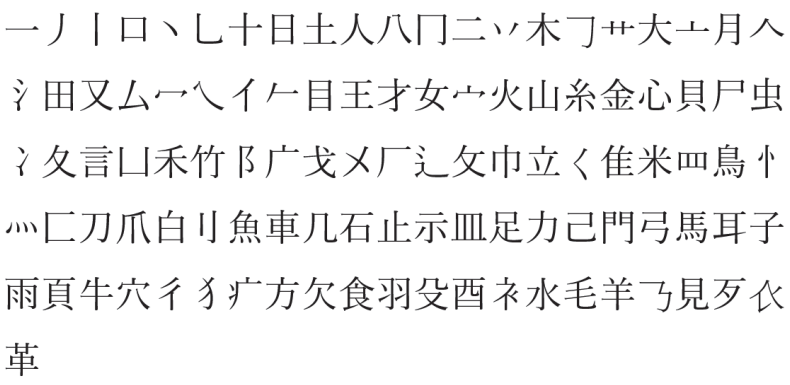

図5　106個の部品リスト（出現回数の多い順）

4-4　部品の同定

例えば「こざと（防・阪の左側部品）」と「おおざと（郎・郵の右側部品）」といった見た目上全く同じ部品を区別することは検索において非効率であると考えられる。また「己・已・巳」といった区別の煩雑な部品も1つにまとめて指定できた方が検索しやすい。さらにUnicodeの10万字を検索対象とする際に、中国大陸の簡体字部品を非簡化部品と区別して検索できる必要もないと考えた。そこで、いくつかの部品同士を同定して検索するためのテーブルを用意することとした。具体的な同定部品テーブルの一部を図6に示す。全体で69レコードある。テーブルの中で全く同じ部品が2つ並んでいるように見える箇所（「舟」「阝」「飠」）があるが、これは漢字としての部品と、「康熙部首ブロック（U+2F00～U+2FDF）」による部品とを同定しているためである。

臣	匝	臣		
舟	舟			
⺾	⺾	⺾		
西	覀	覀		
見	见			
言	訁	讠		
貝	贝			
足	足			
車	车			
辶	辶			
金	釒	钅		
門	门			
阝	阝	阝		
青	靑			
頁	页			
食	飠	食	飡	饣
馬	马			
魚	鱼			
鳥	鸟			
黄	黃			
黒	黑			

図6　同定部品テーブル（一部）

4-5　ツールの実装

以上をまとめて入力補助ツールを制作した。完成したアプリケーションを起動した画面が図 7 である。

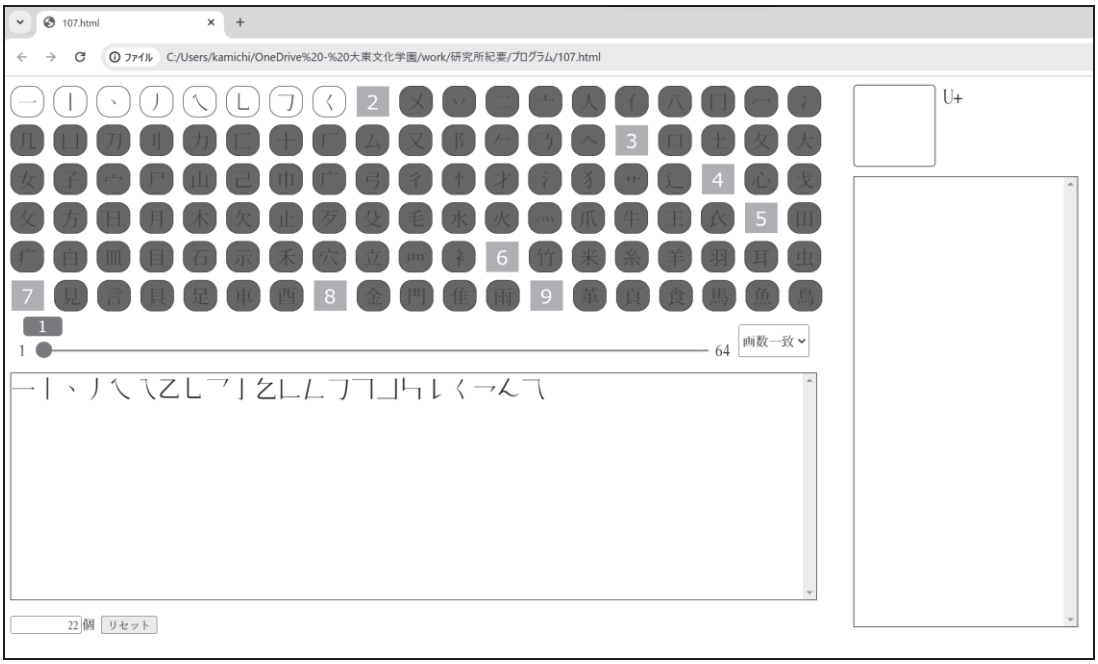

図 7　初期状態（総画数 1 画の検索結果に相当）

左上に並んでいるものが 106 種類の部品ボタンであり、画数順に並んでいる。その下にある範囲スライダは画数を絞り込むつまみである。スライダの右には±の範囲指定ができるようになっていて、スライダの画数に厳密に一致する検索、±1 画の幅を以て検索する検索、±2 画、±3 画を指定できる。このことにより、国・地域によって画数の異なる漢字検索に対応させている。

部品ボタンをクリックすると、その部品を含む漢字に絞り込まれる。グレーアウトしているボタンは、現状で絞り込まれた対象に含まれない部品を指す。図 7 は、総画数 1 画に一致する漢字を検索した初期状態で、下の四角に 22 個の検索結果が並んでいる。

図 8 は総画数 12 画に一致する「土」「皿」を含む漢字の検索結果である。14 個の漢字が表示されている。ピンク色のボタンが絞り込みに指定した部品を指す。白い部品ボタンをクリックするとさらにその部品で絞り込むことになる。

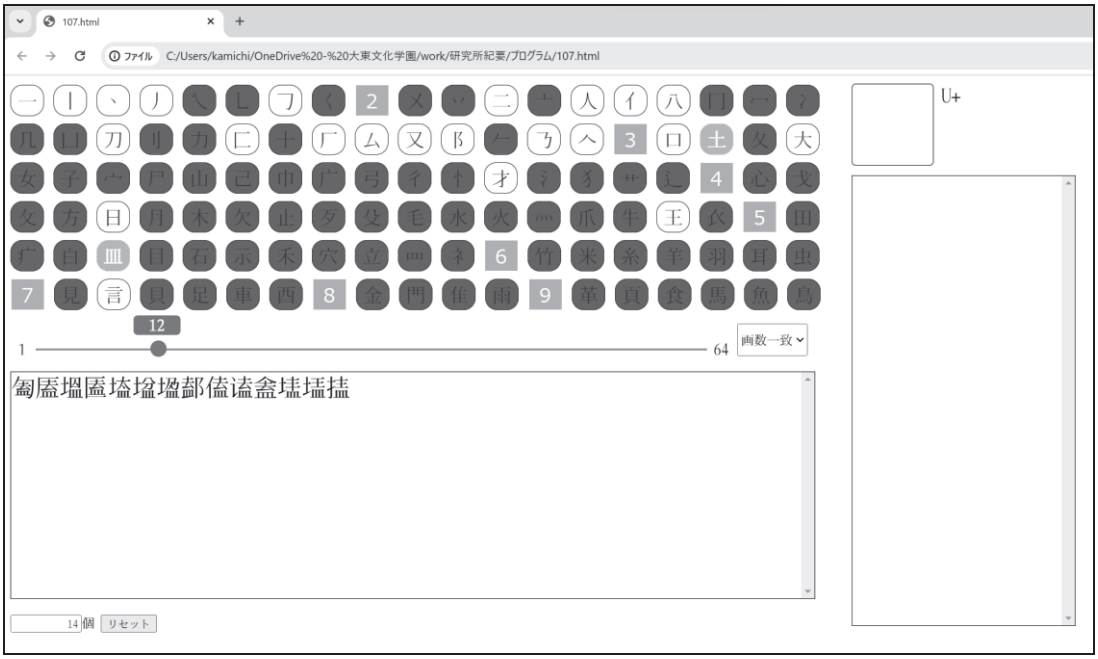

図８　検索結果その１

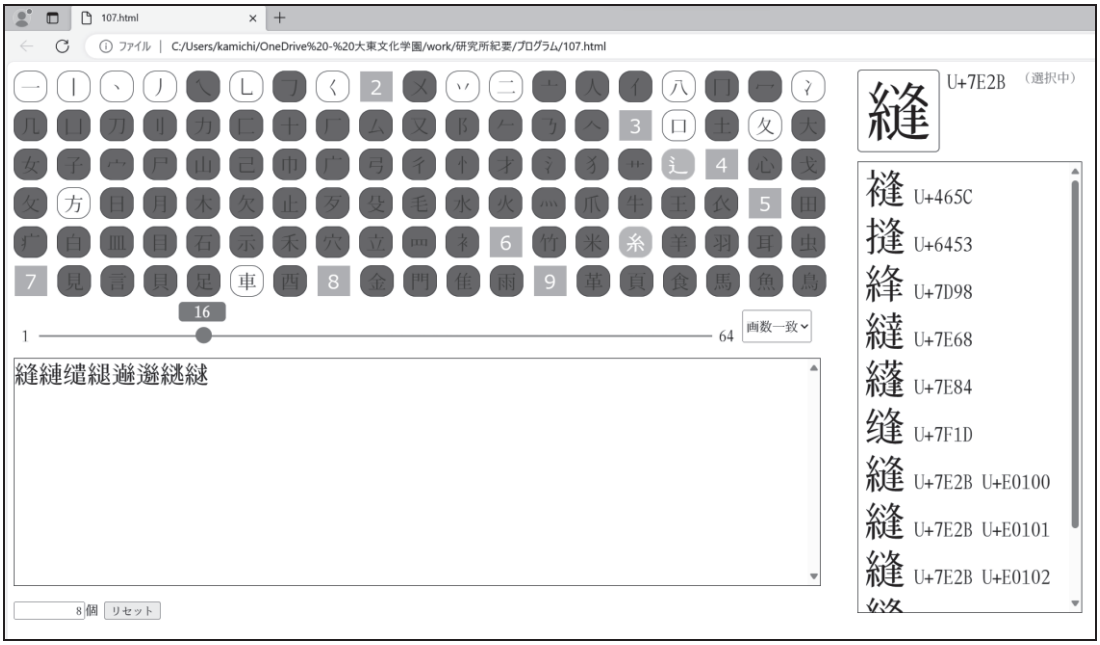

図９　検索結果その２（関連字および IVD の表示）

図９は総画数 16 画で「しんにょう」と「糸」を含む漢字の検索結果である。検索結果の漢字をクリックすると、画面右部分にその漢字が拡大表示され、コードポイントが表示されるほか、グリフウィ

キに登録されている異体字関係にある別のコードポイントの漢字が列挙される。また IVD に登録されている、選択した漢字をベースとする異体字グリフも併せて表示される。図 10 はその拡大図であるが、検索結果の中の「縫」をクリックしていて、上の 6 字がその異体字関係にある別漢字であることと、下の 5 字（そのうち 2 字は下に隠れている）が「縫」の異体字（IVD）グリフ（1 文字目は 1 点しんにょう＋夂の末筆がトメ、2 文字目は 2 点しんにょう＋夂の末筆が払い、3 文字目は 1 点しんにょう＋夂の末筆が払い…）であることを示している。これらの文字を他のアプリケーションにコピーアンドペーストすることで漢字を入力する補助となる。このツールによって Unicode の 10 万の漢字と 3 万の異体字を余すところなくパソコンに入力することができる。

図 10　図 9 の一部拡大

５．おわりに

5–1　現状における課題

本稿では大規模漢字集合に対応したフリーフォントの制作と、入力補助ツールの制作を行った。現状で判明している問題点は主に 2 つである。

① OS・アプリケーションにおける文字コードの対応のタイムラグ：セキュリティ対策から、OS やアプリケーションにおいて、文字コードで未定義のコードポイントに対しては、特にフォントの指定がない限りフォールバックとしてのフォント表示を行わないとのことである（emk 2023）。つまりせっかく文字コードの改訂で新しい字種が収録され、その改訂に対応したフォントを導入しても、OS やアプリケーション側の対応が遅れると、実際には表示されないケースが発生する。オープンソースであれば自分で修正する・改善要求を出すなどの対応も可能であるが、基本的には待つしかない類の問題である。

② 入力補助ツールの検索スピードの問題：当ツールで最も候補数の多い「総画数 13 画±3 画」で検索を行うと 54,503 個の候補が表示され、それにかかる時間は 7 秒程度（CPU：i7-1065G7、RAM：16GB、OS：Windows 11、ブラウザー：Firefox）であり、快適に操作できるとはいいがたい。こちらも設計思想として 1 ファイル単体で動くこととしているため、これ以上の改善は難しいと思われる。

5-2　将来の展望

今回は画面表示の関係で部品数を 106 個としたが、部品数と検索のしやすさの関係について、漢字スキルの大小がどのように影響するかなど、より詳細な検証を行いたい。これについては稿を改めることとする。

〈注〉

1) たとえば『角川 新字源 改訂新版』（角川書店、2017 年）では親字数を（筆者注：約）13,500 としている。
2) IVD は漢字ブロックではないが同じ表にまとめている。各ブロックは収録後の版において数文字の追加がなされている場合がある。互換漢字ブロックはコードポイントの抜けが 2 つあるほか、一部のコードポイントは互換漢字ではなく統合漢字の扱いとなっている。互換漢字および互換漢字補助はいくつかのサブブロックの集合体となっているため版の記述は省いた。
3) IPAmj 明朝　https://moji.or.jp/mojikiban/font/
4) BabelStoneHan　https://www.babelstone.co.uk/Fonts/Han.html
5) 花園明朝フォント　https://glyphwiki.org/hanazono/
6) 謎乃明朝　https://github.com/ge9/NazonoMincho
7) ISO/IEC JTC1/SC2/WG2/IRG　https://appsrv.cse.cuhk.edu.hk/~irg/
8) Unicode Character Database 内：UnicodeData.txt
https://www.unicode.org/Public/15.1.0/ucd/UnicodeData.txt
9) FontForge　https://fontforge.org/en-US/
10) GitHub 内：kamichikoichi/jigmo　https://github.com/kamichikoichi/jigmo
11) 字雲フォント　https://kamichikoichi.github.io/jigmo/
12) Creative Commons CC0　https://creativecommons.org/publicdomain/zero/1.0/
13) MIT ライセンス原文　https://opensource.org/license/mit/
14) 第 4 章は漢字文献情報処理研究会・第 20 回大会（2018 年 1 月於花園大学）にて筆者が口頭発表した「大規模漢字集合の文字検索を自作する」の内容および発表で得られたご意見をもとに改めて制作しなおしたものである。
15) UniHan データ：Radical-Stroke Index
https://www.unicode.org/Public/UCD/latest/charts/RSIndex.pdf
16) Unihan Radical-Stroke Index　https://unicode.org/charts/unihanrsindex.html
17) CHISE IDS Find　https://www.chise.org/ids-find
18) 文字情報基盤検索システム　https://moji.or.jp/mojikibansearch/basic

19) グリフウィキ手書き検索　https://kurgm.github.io/gwtegaki/
20) http://git.chise.org/gitweb/?p=chise/ids.git;a=tree
21) https://raw.githubusercontent.com/cjkvi/cjkvi-ids/master/ids.txt
22) https://babelstone.co.uk/CJK/IDS.TXT
23) https://glyphwiki.org/wiki/Group:sandbox@496

〈参考文献〉

emk, 2023,「利用者:emk　拡張Ｉの漢字をwebブラウザーで表示する」(https://glyphwiki.org/wiki/User:emk@5, 2024年1月4日閲覧).

日本漢字能力検定協会, 2021,「一番多い同音異義語は？」(https://www.kanken.or.jp/kanken/trivia/category06/16060106.html, Internet Archivers より2021年7月22日時点の記録を2024年1月4日閲覧)

Unicode Inc., 2022, “Ideographic Variation Database, ” (https://unicode.org/ivd/, 2024年1月4日閲覧).

Unicode Inc., 2023a, “The Unicode Standard 15.1.0, ” (https://www.unicode.org/versions/Unicode15.1.0/, 2024年1月4日閲覧).

Unicode Inc., 2023b, “The Unicode® Standard Version 15.1 - Core Specification, Appendix E: Han Unification History, ” (https://www.unicode.org/versions/Unicode15.1.0/appE.pdf, 2024年1月4日閲覧).

Production of Large Kanji Fonts for International Standardized Character Sets

KAMICHI, Koichi

In this paper, I detail the development of a Kanji font containing all of the approximately 100,000 Kanji characters in the Unicode international standard character code released on September 2023, as well as 30,000 variants registered in the ideographic variation database. In addition, I created a tool to search 130,000 characters.

Key words : Unicode, Kanji, Variants, Fonts, Kanji Parts