漢字字形情報管理システムの構築と提案

―日本漢文学研究への応用―

上地　宏一

一　はじめに

いわゆる文系と定義される日本漢文学研究者においても、今日の研究活動においてコンピュータを利用することは珍しくない。検索加工、記録複製、印刷出版、交流発信といったさまざまな研究活動を行う上での道具として、コンピュータをはじめとする情報機器の活用はいわば必須の能力である。

コンピュータで漢字を扱うことは、日常生活における一般的な日本語については過不足無く実現できているが、学術研究としての用途、特に日本漢文学においては、訓読文の入力や漢字の入力についてさまざまな制約がある。そもそも、インターネットに縦書きの文章を公開するのですら満足できない状況にある。

本稿は、特に漢字そのものをコンピュータで処理する際に問題となる「標準的に扱えない漢字」に対する一つの解法となりうる漢字字形情報管理システムについて、構想の提案、システムの構築および評価実験を行った結果を報告するものである。

従来「外字」と呼ばれる、限定されたコンピュータの中でのみ利用者が自由に文字を定義できる仕組みを用いてこの問題を解決してきた。外字はインターネットに代表されるネットワークを介した情報交換には適さないため、新たな解法が模索されてきた。一つは大規模漢字（外字）集合の利用であり、一つはネットワーク上での外字の共有である。

本稿で紹介する管理システムは、この両者を包含するものであり、さらに誰でも利用できる外字データベースという特徴を持つ。ここで述べる「誰でも」とは、「自由―所属や立場によって利用が制限されない」ことと、「簡単―コンピュータに精通していなくても平易に利用できる」という二つの意味を持つ。

二松学舎大学二十一世紀ＣＯＥプログラムが運用する日本漢文文献目録データベースにおいて、通常の方法では処理できない漢字に対して管理システムを適用する実験を行ったところ、良好な結果が得られた。システムの提案に至った問題背景の説明を含め、ここに報告することとする。

二　問題の所存は文字コードではない

コンピュータはかつて電子計算機という呼称が使われたように、基本的に数字（数値）のみを処理することができる。さらに、文字と数値とを一対一の関係で結びつけることにより、文章を「数値を列挙したもの」としてコンピュータで扱うことが可能となる。これが文字処理の基本である。それぞれのコンピュータが好き勝手に文字と数値の対応関係を定義していては、コンピュータ同士における情報交換が不可能となるため、工業規格として「情報交換用符号化文字集合」と呼ばれるいわゆる文字コードが各国・地域で制定されている。日本の文字コードは「JIS X 0208:1997」（通称ＪＩＳ漢字コード）である。漢字だけに注目するとＪＩＳ漢字コードには六三五五種類の漢字と数値との関係が規定されている（実際には、ＪＩＳ漢字コードを拡張する形で「JIS X 0213:2004」が新たに規定され、漢字数一万五十字が利用できるが、現状では拡張され

た文字集合が広く活用されているとは言いがたい）。

六三五五種類の漢字には、常用漢字に収録される一九四五字が全て含まれるため、一般的な日本語を表記することが可能であるが、日本漢文学研究者にとって入力できる漢字の種類が足りないということは、少しコンピュータを使うだけで遭遇する問題である。

現在は、国際的に標準化された文字コードであるISO/IEC 10646（一般的にはユニコードという呼称で知られている。厳密には両者は異なるものであるが、文字と数値の対応関係は同じである）においては、日本だけでなく中国、台湾、香港、韓国、北朝鮮、ベトナムなどの漢字文化圏を網羅した文字コードが制定されている。これにより数の上では七万字強の漢字がコンピュータで処理できる状況であり、実際に活用している利用者は少ないと思われるが、マイクロソフト社のウィンドウズ・ビスタには、この七万字の漢字を利用できる環境が整っている。さらにISO/IEC 10646を審議する委員会では、現在も収録要求のある文字集合に対する審議を継続していて、今後も収録漢字数が増える方向にある。

しかしながら文字コードの拡充が諸問題を解決するものでは決してない。その理由として文字コードが漢字の字形を規定するものではなく、言語情報を交換するための文字の概念を規定するためのものであることに起因する。具体的には、人間が区別する字形の差を文字コードでは区別しない場合がある。ＪＩＳ漢字コードでは「この規格は、字体の図形的実現としての字形については規定しない。一つの字体の図形的実現としては、デザインの差に基づく複数の字形が考えられるが、この規格はそれらを互いに区別しない」とある。また、ISO/IEC 10646国際規格を日本国内で整合させた規格であるJIS X 0221-1:2001では「図形記号は、文字の代表的な可視化表現とみなさなければならない。この規格群は、各文字の形を正確に規定しようとするものではない」と明記されている。具体的な例を挙げると「珊」と「珊」は同じ番号（ユニコード番号73CA）が与えられているため、テキストデータとしては区別することができない。特に漢字を専門に扱う研究者においては、このような状況は大変不便であるが、細かい字形の区別は文字コードの範疇ではない。

また、文献から全くの新しい漢字が発見される場合もある。日本漢文学に関する古典籍のデジタル化は『四庫全書』や『四部叢刊』に代表される中国の古典籍と比較して量的に途上にあることからも、今後新たに数十字単位で発見される可能性もある。新字種については先述のように文字コードに追加を要求することが可能である。しかしながら無尽蔵に要求を受け付けるのではなく、文字として定まっていることの証明としてまとまった典拠情報の提出が求められ、さらに文字の追加には長い審議期間が必要とされる。ISO/IEC 10646 における漢字集合の追加はＩＲＧと呼ばれるＩＳＯの中の部会で審議されるが、そこでは日本だけでなく中国・台湾・香港・マカオ、韓国・北朝鮮、ベトナム、アメリカ、ユニコードなどのメンバーが参加し、おのおのが追加漢字集合を申請する。あらかじめ想定していた追加枠に収まらない数量の申請があることや、メンバー全体による申請文字のレビューによって採録が認められないこともあり、申請した文字が全てそのまま収録されるとは限らない。その後正式に収録されたとしても、実際にパソコンでその追加文字が処理できるようになり、フォントが普及するためにはさらに長い年月が必要となるため、今日明日この文字が使いたい、という要求には現実的に到底答えられるものではない。このように「標準で使えない漢字」を文字コードの拡張によって解決することは難しい。

三　既存の対処方法

そこで従来は主に外字（ユーザー定義文字）を利用した解決がなされたほか、標準規格以外の文字コードを活用した方法などが提案されてきた。本稿で提案するシステムと比較対象となる製品・技術標準および先行研究について、以下にいくつかを挙げる。

（一）今昔文字鏡、インデックスフォント

第一に挙げられるのが、ＪＩＳ漢字コードの枠組みを利用しつつ、フォントを切り替える方式によって十五万字形が処理可能となる製品である「今昔文字鏡」である。製品の内容は、十五万字形を納めたフォントファイルと、その字形を検索するための検索プログラムである。フォントは諸橋轍次『大漢和辞典』の見出し字を丸ごと収録しているため、日本の文学研究者に受け入れられた結果、学術研究機関等での導入も見られるが、フォント字形や漢字字形に付番される文字番号の利用に許諾申請が必要なため、最近では使用が敬遠されるケースもある。これに対して、出版や商業印刷で利用可能なライセンスが付与される「インデックスフォント」も販売されている。通常のフォントとは文字配列の異なるフォントを利用するため、デジタルテキストデータの閲覧者はインターネット上で無償配布されているフォントをあらかじめ入手する必要がある。フォントは通常の明朝体とは異なる独特の文字デザインが施されているため、通常の明朝体本文に埋め込んだ際、見た目に違和感を覚える人もいる。

文字の追加申請は製品購入者および文字鏡研究会会員が可能で、一回の請求で二十字以内、二ヶ月の制作期間となっている。追加されたフォントの無償一般公開分への反映は不定期とされている。

（二）超漢字（トロンＯＳ）

現在普及しているウィンドウズとは別のＯＳとして多漢字処理を実現するのがトロンと呼ばれるＯＳの仕様に準拠した「超漢字」という名前の製品である。当初日本のＪＩＳ漢字コード、および中国の国家規格文字コードであるGB 2312、韓国の国家規格文字コードであるKS X 1001の三つを共存する形で設計された独自の文字コード（トロンコード）を利用している。トロンコードはその後、台湾や香港で普及しているBig5や、今昔文字鏡（その後製品からは削除された）、大漢和辞典見出し字などが追加され、さらにＧＴ書体と呼ばれる大規模文字集合が加わった。ＧＴ書体そのものはウィンドウズで

も利用できるフォントであり、無償配布されている。

トロンコードおよびＴＡＤと呼ばれる独自のデータ形式は他のＯＳとの情報交換において変換作業を必要とするため、研究成果の対外公開が求められる今日において利用するメリットは小さく、ＯＳの普及度も低い。また、ＧＴ書体（トロンコード）を用いたテキストデータもインターネット上で公開されているものはほとんど見受けられない。

トロンコードに未収録の文字については収録申請を行うことができる。収録にかかる期間についての説明は見受けられないが、二〇〇一年から二〇〇七年の間に合計で九回の収録文字集合更新の告知がなされている。また、「超漢字」開発元のパーソナルメディア社では独自の文字追加を含めた多漢字処理システムの提供を行っているが、対象は自治体などであり、費用や事業規模の説明から判断すると文学研究者・研究機関を対象としているとは考えにくい。

（三）ＸＫＰ

ウィンドウズＮＴ拡張漢字処理（略してＸＫＰ）とは、情報システムメーカーが提供する外字集合をある一定のネットワーク内のコンピュータで共有・管理する処理システムの技術仕様[1]を指し、人名・地名外字処理が必要な企業や自治体から注目された。ＸＫＰは技術協議会が発足した一九九五年当時の状況として、メーカー独自の文字コードや、ＪＩＳ漢字コードで構築されてきた既存システムにおける外字集合やデータベースをウィンドウズの内部文字コードに新たに採用されたユニコードに対応させることを目的としたものであり、現在は直接活用する機会は少ないが、ネットワークで外字を共有するユニコードを基盤とした処理システムの例として紹介すべき技術である。

（四）戸籍統一文字

戸籍のデジタル処理のために平成十六年に法務省が定めた[2]文字集合で五万余字からなる番号表であり、戸籍手続きのオン

ライン化のために戸籍に利用できる文字集合を整理したものである。この文字集合を各自治体が共有することによりシステム（メーカー）の異なる自治体間でのデータ処理が可能となっている。当然のことながら人名の異体字に特化した文字集合であり、文学研究者の要求を満たすものではない。また、一般に利用できるフォントが提供されているわけではない。

（五）ISO/IEC 10036

文字字形（グリフ）を登録し一意の番号を付与する機構をもった国際標準規格である。利用者は手続きを行うことで登記簿に自由に文字字形を登録することが可能で、各字形に対して付与された番号は構造記述が可能なＸＭＬ文書において活用し、文書の閲覧者がその文字の具体的な文字字形を参照することを可能とするものである（ここの文字はISO/IEC 10036のこの番号の文字字形に該当する、といった具合である）。登録簿の一部はインターネットで公開されていて自由に閲覧できる。先述の今昔文字鏡で公開されているフォントに含まれる文字字形はこのISO/IEC 10036に登録されている。

大きな特徴として文字字形を登録する際に既存登録分との重複のチェックを行わない方針が採られている。これは重複しているかどうかの判断は人の主観によって左右されるため、としている[3]。このほか文字字形以外の情報は登録しないため、ユーザーがテキストを入力するときにISO/IEC 10036の文字字形から自分の意図する文字を探す、といった活用はできない。また文字字形の登録には登録料が必要である。これは登録の濫用を防ぐためでもあるとしている。

現状ではISO/IEC 10036が広く活用されているとは言えないが、今後字形識別の手段として普及する可能性もある。

（六）ユニコード漢字異体字データベース（ＩＶＤ）

デジタル文書内で字形を書き分ける方法として、文字に続けて字形番号に相当する記号を記述する方法がユニコードの技術文書として規定されている[4]。字形番号は個々のユーザーが勝手に使用するのではなく、ユニコードコンソーシアムに所定

の方法で登録申請を行い、データベースに登録されたものだけが利用できる。現在は各文字に対して二百四十個の識別記号が用意されているが、要請に応じた記号数の拡張についても言及されている。現時点でデータベースはまだ存在しないが、実質的に広く普及している文字コードであるユニコードにおいて正式に規定された字形指定方式であり、将来ＯＳやアプリケーションで標準的に利用できる仕組みが提供される可能性が高い。

現在登録申請手続きが行われているのは、ＤＴＰや画像・映像技術ソフトウェアメーカー大手のアドビ社が定義する、異字形を多く含む独自の日本文字字形集合（Adobe-Japan1）の登録であり、アドビ社はこの文字字形集合で記述されたデータを標準的なデータとして利用可能とするために今後積極的にこの技術を活用するのではないかと推測される。データベースへの申請方法などを勘案すると、ＩＶＤはアドビ社やフォントメーカーのようなソフトウェア・フォント供給団体が自己所有する外字や文字字形集合を登録する手段であり、個々のユーザーや研究機関が活用することは考えにくい。

（七）日本語外字センターの提唱

小山氏が提唱する[5]日本語外字センターとは、住民基本台帳ネットワークの人名処理において外字となる漢字について、各自治体から照会された字形情報を一元管理し、フォントの配信や文字情報の照会を可能とする機関を設置するという構想である。先述の戸籍統一文字の制定により、外字処理が必要となる文字は原則的に無くなったため、このセンターの必要性も無くなったのであるが、センターが字形情報を一元管理するということは、各自治体・メーカーごとに分かれていた字形情報の同定作業をセンターが主体として行うことであり、「統一された思想による文字種の判定」を行うことである。そもそもある字形とある字形の同定作業が可能であるかどうかという問いに対して「誰もが納得できる字形の判定方法、及び標準化は不可能であるという結論に達しました」と踏まえたうえで「文字コードの束縛を受けない、ほぼ無限の字種字形を包含可能なファイル名（筆者注：漢字字形に対して付与する一意の名称）を利用し、判定を日本語外字センターという一組織に

任せるという方法にたどり着いた」と述べている。その方法として「状況をインターネットにより常時情報公開すること」で利用者が評価することにより恣意的な判断を防ぐことができる、としている。このように、文字字形に名前を付与して識別を行うこと、およびインターネットによる衆目監視が外字管理（異字形の同定）に有効ではないかという二つの意見は注目すべきである。

以上のように、標準的な文字コードの範疇を超えた漢字情報処理が必要な自治体や企業のニーズを反映した外字処理システムや、文学研究者も活用する大規模外字フォントなどによりある一定の解決が図られてきた。

四　ユーザー本位である外字管理環境の欠如

ところで日本漢文学研究者の中でもパソコンに興味を持っていた人の中には、ＮＥＣ社のパソコンであるPC-9800シリーズを活用していた人も多いのではないだろうか。その頃は文字コードに不足する外字については各ユーザーが自分で作字して管理していたケースが多く、筆者もいくつかの外字フォントを所有していたことを記憶している。それが現在では他者が提供する外字フォントを使うなどの受動的な立場に移行せざるを得なくなった要因については次の二つが考えられる。

（一）インターネットの普及に代表されるコンピュータ同士の情報交換に対する外字利用の弊害
（二）漢字の外字フォント作成やフォントの配布が面倒である

これらの要因は密接に関連している。すなわち、文字字形のデータ形式で現在主流のアウトライン形式は、表現豊かで精

細な印刷が可能である反面、文字字形のデザインおいて習熟した技術を要するものであり、複数のフォント編集ソフトウェアが提供されているが、フォントデザインの非専門家にとっては敷居の高いものとなっている。また、既存の漢字フォントに含まれている漢字部品を持ち合わせて外字フォントを作成することも考えられるが、一部のフォントでは使用許諾契約（ソフトウェアライセンス）によって、フォントデータの加工や複製といった二次使用に制限が設けられているため、インターネットでの外字フォントの公開といった配布・複製行為が行えないケースもある。このように、外字フォントの作成や公開が困難であるため、外字利用の必要な古典籍に関する研究分野において将来的に公開の可能性があるデジタルテキストを作成するには、フォントの共有可能性の高い外字フォントの利用が必須となり、必然的にユーザー固有の外字フォントの利用が廃れていくと考えられる。多くの場合、既存の外字フォントを利用することでニーズを満たせるはずであるが、それでも足りない漢字字形というのはゼロにはならないと考えられる。

このような外字の利用において受動的な立場にある現状を鑑みると、必要な漢字字形を自由に・容易に利用できて、かつ情報交換に支障のない、漢字字形情報管理環境が切望されているのではないかと考えるに至った。

五　漢字字形情報管理システムの提案

そこで能動的な外字管理環境の構築を目的として、先述の二つの要因を解消する漢字字形情報管理システムを提案する。この管理システムはいわゆるウェブデータベースの形態をとり、データ管理をインターネット上で行う。つまり、外字データの登録や編集をインターネット上で行い、情報の更新がすぐに公開内容に反映されるようにする。このことで、デジタルテキストを作成する際、足りない漢字（字形）があれば、管理システムの外字データを検索し、既に用意されている場合はそれを用い、登録されていない場合はその場で登録を行うことができる。別のユーザーがデジタルテキストを閲覧するため

に外字フォントが必要である状況は今までと変わらないが、必要な外字フォントは、システムから自由にダウンロードできることとする。このため、データ作成者は外字フォントを公開・配布する作業から解放される。もしくは、登録されている字形を画像ファイルとして取得しテキストに埋め込む（または、ＨＴＭＬドキュメントから画像ファイルを参照する）利用方法も考えられる。

データベースに外字データを登録する際に、個々の漢字字形は一意の名前を付与することで識別を行い、命名規則は特に設けない。また登録する文字が既存のデータに含まれているかどうかの判断は各ユーザーが行うこととし、同定判断の審査は行わない。その代わりに漢和字典や文字コードにおいて規定されている文字、出典が明確な文字などについては（例として「諸橋大漢和二七五一番」「ISO/IEC 10646 の U-00020000」など）、強制力のない命名ガイドラインを設けることとするが、ユーザーが判断できない、または意図的に独自の外字集合としたい場合は、ユーザーが自由に命名することになる。さらに、ユーザーごとに操作できるデータベースを区別せず、全てのユーザーが全てのデータにアクセスできる特徴を持つ。これにより他のユーザーが作成したデータをそのまま、あるいは部分的に活用することが可能となり、同時に自分のデータを他人に提供することにもなる。このようにデータの内容についての管理者がいないデータベースであるため、同名による漢字字形が登録された場合は、新しいデータによって上書きすることとなる。ただし、過去の全履歴を参照可能で、版の指定により個々の版のデータを一様に利用できることで、登録された全データの利用を保証する。また、各ユーザーが作成・編集する権利を占有するデータを登録するための命名規則を特別に用意することで、ユーザー独自の外字集合を登録することを可能とする。ただし、この場合でも登録されたデータの参照や二次利用の権利は一般に開放される。

データの管理者を置かず、文字の登録に恣意的な判断をせず自由に登録できることを定義した結果、漢字字形登録システムを実際に設置する上で、設置者の選定（私企業が行ってもよいのか、中立的な国家機関が行うべきか、そもそも国が文字字形の判断を行ってよいのか）に迷うことも無く、また大掛りな予算が不要で、ビジネス面での需要と供給を満たさなくて

も設置できるため、小さな外字集合の利用を目的とした研究者などの個人ニーズをも満たすことが可能となる。

（一）ウィキシステムの利用

このようなウェブデータベースを実現するために、ワード・カニングハムが提唱したWikiWikiWeb[6]（現在では略してウィキと呼ぶことが一般的であり、本稿ではこの概念を実現したウェブシステムのことをウィキシステムと記す）をシステムのベースとした。ウィキシステムとは、元来情報の提供者と閲覧者が明確に分離されているインターネット（ウェブ）の世界において、掲載されているテキスト情報（ウェブコンテンツ）を閲覧者が編集（新規作成、更新、削除）できる仕組みであり、ウェブ上において複数人が共同で推敲などの作業を行うことを可能としたものである。ウィキで扱うことのできるテキスト情報は、文字だけではなく一般のウェブコンテンツと同様に文字の装飾、画像の埋め込みや他のページへのリンク（ハイパーリンク）も扱うことができる。ウィキシステムにおけるコンテンツの記述はいわゆるＨＴＭＬの書式と比較するとより平易な独自の書式が用いられており、このことからウィキシステムを管理コストの小さいコンテンツ管理システムの目的で用いた利用形態も散見される。これはブログと呼ばれるウェブ日記システムが、日記本文を打ち込むだけで自動的にウェブコンテンツとして加工され、インターネット上に公開できる便利さと類似する。

ウィキシステムの中でも特に一般に知られているのが「フリー百科事典」であるウィキペディア[7]（Wikipedia）だろう。イギリスの科学雑誌『ネイチャー』が、ブリタニカ百科事典とウィキペディアの内容比較記事を報じる[8]ほどであり、英語版では約二百万項目、日本語版においても約四十三万項目の記事が投稿されている。これらは全ての記事が一般ユーザーに開放されたウィキシステム上で構築されたものであり、ボランティアによって日々内容の拡充が続けられている。記事内容の信頼性・中立性や分野ごとの記事の偏りといったさまざまな問題点が存在するが、ウィキという新しい枠組みでこれだけの発展が見られることは、バザールモデルの成功事例として無視することはできない。バザールモデルとは、エリック・レイ

モンドが発表した『伽藍とバザール』[9]と題する著書において主張しているコンピュータソフトウェアの開発の方法論であり、ソフトウェアの設計図にあたるプログラムソースを常に公開し、ボランティアのプログラマーに開かれた開発環境を用意することでネットワーク上の多くのプログラマーが参加し、よりよいソフトウェアを構築することを可能とするものである。ウィキシステムはこのモデルを応用したものであり、インターネット上の沢山のユーザーを集めて大きな作業を行うという手法は新しいウェブ利用形態の一つであり、現在注目されているモデルである。

漢字字形情報管理システムでは、このウィキシステムの二つの特徴である「共同・協同作業としてのツール」「操作が平易なウェブシステム」を応用し、外字管理データベースシステムを構築するものである。また、副次的な狙いとして、フリー（自由・無償）なフォントの協同制作のツールとしての活用も考えている。すなわち、ISO/IEC 10646 において定義されている七万字のうち、五万数千字については日本のフォントとしては提供されていない（先述の通りマイクロソフト社のウィンドウズ・ビスタでは中国大陸デザインおよび台湾・香港デザインの二種類の七万字フォントが用意されているが、制作コストと実際のニーズを比べると今後日本デザインでの七万字フォントが提供されることは考えにくい）。この空白部分について、ネットワーク上の一般ユーザーが分担して漢字グリフを埋めていけば、少ない制作コストで、しかもフリーなフォントができるのではないだろうか。これはバザールモデルを用いた大規模漢字フォントデザイン、つまりウィキペディア百科事典ならぬ漢字グリフ百科字典の実現である。当然のことながらデザイン品質や字体の統一性など、考えられる問題点は複数あるが、従来にないフォントデザインのモデルとして思案中である。

（二）ＫＡＧＥシステムの利用

漢字字形情報管理システムの構築において、大きな問題となるのが「平易な漢字字形デザイン手段」の確保である。文字字形の輪郭線を直線とベジエまたはスプライン曲線の集合で表現する既存のアウトラインデータは、デザイン技術の習得が

困難である。また輪郭線による字形データは、線形的な拡大や縮小を施した場合に、筆画の太さに対しても変形がかかってしまうため、漢字部品を組み合わせて文字のデザインを行う際、太さの修正が必要となる。アウトラインデータを用いることで既存の活字字母の再現を含めた高品質な文字デザインが可能となるが、研究者にとってはとにかく文字（字形）が印刷できれば良い、と考える場合も多い。

そこで、管理システムでは、既に開発していたＫＡＧＥシステム[10]を利用することとした。ＫＡＧＥシステムとは、漢字字形を筆画ごとの骨格情報の集積で表現する、グリフデータの中間表現形式である。現在フォント処理で主流のアウトライン形式によるデータの最小単位は、輪郭を構成する一つの座標点であるが、ＫＡＧＥシステムでは漢字字形を筆画に分解したものをデータの一単位とする。各筆画は線の種類（直線、曲線など）を指定し、筆画の骨格に相当する中心線の位置情報および筆画の形状を情報として記述する。中心線は、直線の筆画の場合には始点と終点の座標情報を、曲線の筆画はさらに一、二点の制御点の座標情報を加えた、二点から四点の座標情報で表現する。筆画の形状は、筆の入り、跳ね、他の筆画への接続、止め、といった数種類の形状を頭部と尾部の二種類について記述する。これら数項目の情報で一つの筆画をあらわし、筆画種数分のデータを組み合わせて一つの漢字字形データを構築する。このデータを元に、明朝体の特徴を持つアウトライン形式のデータへの変換プログラムを用意することで、漢字字形の画像ファイルや一般的なフォントファイルに変換することを可能とする。また筆画による表現だけでなく、漢字部品を位置と大きさの二つの情報の記述により引用することも可能であり、漢字部品を組み合わせて新しい漢字字形を容易にデザインすることが可能となっている。例として「永」という漢字字形を表現するためにアウトライン形式では六十二単位の情報が必要であるのに対し、ＫＡＧＥシステムでは三十九単位で表現が可能であり、「議」では百八十七単位に対して八単位での表現が可能となっている（これほど少ない情報量であるのは「言」と「義」の部品引用を行っているためであり、筆画ごとに分解した場合は百十二単位の情報量となる）。

ＫＡＧＥシステムの漢字字形データは数値の集合で表され、これを一行一筆画に相当するテキストデータとして表記する

ことができる。このため、先述のウィキシステムの一記事分を当管理システムでは漢字一字形と見なし、ＫＡＧＥシステムによって表現された漢字字形データを一つの記事として登録する方式を取ることで、データベースの設計はウィキシステムの仕様をそのまま利用することが可能となった。実際にはＫＡＧＥシステムによる漢字字形データを数値の記述でデザインすることは難しいため、座標点をマウスで操作することでデザイン可能なグリフエディタを用意している。このエディタは一般に普及しているアドビ社のフラッシュコンテンツとして実装しているため、管理システムを操作するウェブブラウザ上でそのまま利用し、またデザイン編集した結果を同じウィンドウ内で管理システムに登録できる。

（三）データライセンスの決定

管理システムでは、漢字字形データを自由に登録し、自由に活用することを想定している。そのため、著作権などに制限されないデータのライセンスを決定する必要がある。ウィキシステムとして参考としたウィキペディアでは、投稿する記事のライセンスはＧＦＤＬ[11]というフリーソフトウェア財団（ＦＳＦ）が提唱している文書を対象としたライセンスにウィキペディアの独自の解釈を加えたものが適用されている。ＧＦＤＬライセンスの根底にあるのはＦＳＦが提唱する「コピーレフト」という概念で、フリーとは、「無償」ではなく「自由」を指し、将来にわたってそのデータおよびその全ての派生物が自由に利用できる状態を維持できることを保証するための制限が設けられる。このためコピーレフトを主張するデータを、別のデータと結合して新しいデータを作成した場合、新しいデータ全体がコピーレフトとなる必要がある。漢字字形情報の場合、漢字グリフに著作権が発生するかどうかは、少なくとも日本の現行法においてフォントに著作権が存在しないという見方が強いことからも否定的に解釈するべきであるが、仮に著作権が存在すると見なしたときに、その漢字グリフを集めてフォントファイルに変換し、そのフォントの漢字グリフを埋め込んだ文書ファイル（ＰＤＦ形式などが想定される）を作成した場合には、他の著作物と合体することになる。その際に漢字グリフにＧＦＤＬライセンスを適用しているとフォントフ

ァイルとして合体した著作物（文書ファイル）全体に対してＧＦＤＬライセンスの適用が求められる可能性が生じる。つまり、作った文書ファイルに対して他者による自由利用の保証が求められることになるが、これは現実的ではない。もっともＦＳＦが別に提唱している、ソフトウェアを対象とするＧＰＬ[12]ライセンスでは、フォントファイルにライセンスを適用した際の文書へのフォントの埋め込みについて、埋め込んだ文書に対してはＧＰＬライセンスの適用を行わない旨の特記事項を付記することでＧＰＬライセンスをフォントに適用することも可能である、という解釈がなされている[13]。いずれにしても「コピーレフト」の思想は、将来にわたる自由利用を担保するために、著作物の利用者においてライセンスに対する一定の注意が求められることになり、管理システムの求める自由利用の方向性とは一致しないと考えた。そこで、管理システムとして一次的にのみ利用の自由を保証し、そのデータを二次利用して作成した派生著作物に対しては自由の保証を求めない最も緩い形式のライセンスとすることにした。その方法としては、ユーザーがデータを管理システムに登録する際、そのデータはシステムを運用する主体者に権利を移譲することとし、さらに日本の著作権法では放棄できないとされている人格権については、今後権利を行使しないことを了承してもらった上でデータの登録を受け付けることとした。そのデータを管理システムでは条件をつけずに自由な利用を認める形で公開する。このことによりシステムに登録するデータは将来的にいかなる形態の利用も妨げないことを保証できる。権利の移譲については、他のブログサイト等にも見られる方式であり、著作権権利の完全放棄が法的に有効かどうか、移譲契約として有効かどうかの厳密な判断は不明であるが、運用ポリシーの明示という点においてユーザーに注意を促すことで自由利用の保証に関する理解を得ることができるものと考える。

（四）ウィキシステムの構築

先述の通り、ユーザーが必要な漢字字形を自由に登録し、自由に利用できる漢字字形管理システムを提案し、実際のウィキシステムを構築した。その際、ウィキペディアの運用システムである MediaWiki というシステム（ソフトウェア）を参

考とした。もともとウィキシステムは不特定多数のユーザーによるデータの編集がなされるため、データの統一性保持や運用に一定の権限を持った管理者の存在が望まれる。しかしウィキペディアのように参加するユーザーが多く、特定の管理者に頼っては運用できない場合は、データ細部の調整自体もユーザーが行うことが要求される。このため MediaWiki では「ノート」と呼ばれる、個別の記事に対してユーザーがコメントを記述できる場が用意されている。例えば主観の異なるユーザー同士が一つの記事に対する差し戻し編集を繰り返すような場合に、意見調整の場としてノートが活用される。そこでは、当事者の議論だけでなく傍観者からの賛否の表明も行われ、一定のルール（意見表明期間の設定と多数決による決議）に則り解決が図られる。必ずしも全ての議論に対してノートが有効であるとは言えないが、ウィキシステムの中でも特徴的な機能である。

漢字字形管理システムの場合、想定される利用ユーザー数はそれほど多くないが、データ内容についての管理者は不在であり、また文字のデザインは見慣れているかどうかなどの主観が大きい分野であるため、このような意見の調整機能が有効であると考え、この「ノート」機能を取り入れることとした。このほかウィキペディアそのものが、ウィキシステムの中でも良く知られ実際に活用している人も多いと考え、管理システムの構築に際して MediaWiki の仕様をそのまま用いることとした。ただし登録データの対象が漢字グリフでありソフトウェアの機能改造が必要な工数が多いと考え、ソースが公開されている MediaWiki そのものは用いずに、仕様だけを参考に独自にソフトウェアの実装を行った。管理システムの名称は「グリフウィキ」とした（末頁の図参照）。

六　日本漢文学への応用

（一）目録データベースへの適用

二松学舎大学二十一世紀ＣＯＥプログラムでは日本漢文学に関する資料や論文の書誌情報を「日本漢文文献目録データベース」としてウェブ上で公開している[14]。このデータベースにはフォントが見つからないとしてゲタ文字（〓）扱いになった文字が現時点で約数百種類記録されている。その多くは実際にはISO/IEC 10646において収録されている文字であったが、判定できない、あるいは新字種と考えられる文字がある。これらをグリフウィキに登録することとした。そして登録したグリフを目録データベースにおいてゲタ文字の代わりに画像ファイルとして表示する実験を行った。

グリフウィキでは、グリフデータを登録すると直ちに画像ファイルが生成され、ウェブを通じて利用できる状態となる。その画像ファイルは二百ドットあるいは五十ドットの大きさであり、実験では五十ドットの画像を通常の文字の中に埋め込む形式を採った。目録データベースではゲタ文字に続けて外字番号を入力しているため、外字番号をそのままグリフ名に引用し、外字グリフをグリフウィキに登録することで、データベースの検索フロントエンドプログラム内で機械的にゲタ文字をグリフ画像に置き換えることが可能となっている。

実験の結果、五十ドットの外字画像と十六ドットや十二ドットの通常の文字との差が大きいために見た目の違和感はあるものの、ゲタ文字ではなく意図する文字が再現できている点で目的は達成しているものと評価できた。課題としては、このようなウェブコンテンツからのグリフ画像の呼び出しに対応するため、より小さなグリフ画像のファイル生成に対応すべきであることがわかった。

（二）版本表現型全文テキスト表示への適用

目録データベースと同様にＣＯＥプログラムで実施している「儒蔵プロジェクト」で検討中の版本表現型テキスト表示プログラムにおいて、外字扱いとなる文字をグリフウィキのグリフ画像ファイルで置き換える実験を行った。版本表現型テキスト表示とは、全文テキストの元となる版本の文字の配列を忠実に再現したテキスト表示のことである。置き換えの対象

は、全文テキストの入力を行った竹添光鴻『左氏会箋』とした。版本表現型テキスト表示の実装における出力テキストのデータ形式については現在も検討中の段階であるが、ここでは案として挙がっているＰＤＦ形式での出力を対象に実験を行った。具体的には版本の形式に沿った全文テキストの表示において、外字部分をグリフウィキに登録した画像に置き換えてＰＤＦファイルに出力するものである。ＰＤＦ形式では画像のラスタライズ表示が可能であるため、より高精細な二百ドットのグリフ画像を用いた。

その結果、標準倍率での画面表示において、また六百ＤＰＩのプリンタで印刷した結果において、通常の文字と比較して遜色のない水準での出力が可能であるとの評価に至った（末頁の図参照）。さらに高品質な出力に対応するためには、グリフ画像サイズを二百ドットではなく通常のアウトラインフォントの分解能である千ドット（一〇二四ドット）に引き上げることが考えられる。ＫＡＧＥシステムによるグリフデータをアウトラインデータに変換する際の解像度を引き上げることによって分解能を上げることは理論的に可能であり、高品質出力に対応可能である。また、これだけの高解像度になると、グリフの大きさ（ボディサイズ）や中心線の位置の微妙なずれが逆に目立ってしまうことが判明したため、実際の運用時にはこれらの要素の微調整が必要である。幸い今回対象となる版本表現型テキスト表示の場合、漢字のボディサイズは全ての漢字で同じとなるため、調整は一回でよい。

（三）典拠情報データベースとの連携構想

漢字字形情報管理システムの構想を何度か発表した折、複数の方から、登録されているグリフに対してそのグリフが確かに利用されている典拠情報を自由に登録できると文字データベースの性格を持たせることができてより学術的に価値が出るのではないかというコメントをいただいた。現在のグリフウィキの構想では、登録されている各グリフが持つメタ情報は、「関連字」と呼ばれるISO/IEC 10646に収録される文字との関連付け情報のみである。関連字は、ユーザーの目的とするグ

リフがグリフウィキに登録されているかどうかを判断するときに、ユーザーが関連すると考える一字を入力してもらい、その字およびシステムに内蔵する異体字データベースを参照し導出した異体字関係にある文字群のいずれかと結び付けられている登録グリフを一覧できるようにすることで、重複したグリフを登録させにくくする目的がある。しかしながら異体字ではなく新字種については、関連する一字を指定することは難しく現状ではゲタ文字（関連字なし、の意）を指定することを許容している。

特に今後発見の可能性が高い日本漢文をはじめとする古典籍における新字種をグリフウィキに登録する際に典拠情報を同時に入力し、また既に登録されているグリフが別の出典において用いられている場合にその情報を重ねて登録することにより、その文字の新字種としての確立の度合いが増すことになる。このようにして収集した新字種のデータは新たに文字コードに収録要求を出す際の基礎資料として有益であると考えられる。

このような典拠情報などのメタ情報をグリフウィキでどのように扱うかについては、さらに議論が必要である。その理由としては、典拠情報等を固定した書式で記録する方式が運用可能であるのかどうか、逆に自由書式での情報記述を許容した場合に検索や集計で問題が出るかどうかの検討が不足していると考えるためであり、文字データベースの研究者からのアドバイスを求めるべきである。

例えばグリフウィキはあくまでグリフ情報を登録する手段と考え、典拠情報や他のメタ情報については、ＣＨＩＳＥプロジェクトが提唱・公開しているＣＨＩＳＥ文字情報データベース[15]との連携で実現することも考えられる。ＣＨＩＳＥ文字情報データベースは、文字一つ一つをオブジェクトと見なし、そのオブジェクトに対してさまざまな情報（素性）を付与することでその文字を表現する概念モデルを用いている。一つの文字オブジェクトとグリフウィキの一グリフを結びつけることで、登録されているグリフに対してさまざまなメタ情報を付与することができるものと思われる。

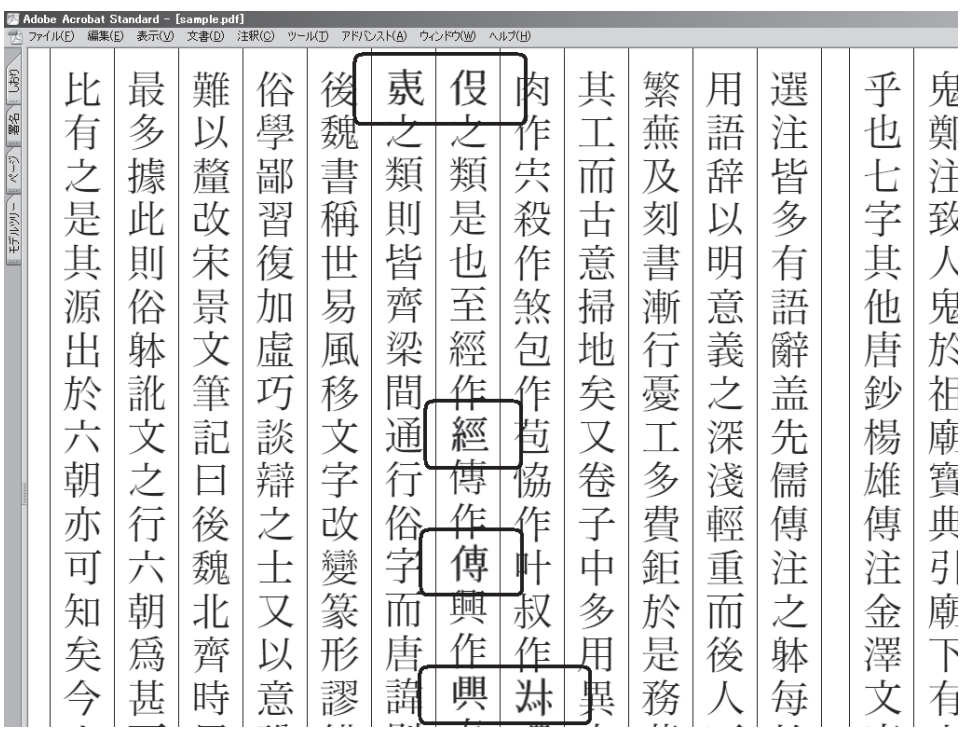
七　おわりに

現在グリフウィキはシステムを公開し[16]、試験運用を実施している段階にある。不特定多数のユーザーによる共同管理やフォント制作が実用的となるかどうか、平易とする操作体系が研究者をはじめとするユーザーに受け入れられるかどうかは今後の検証を待つ必要があるが、少なくとも日本漢文学の資料発信に関する二つの事例において、本システムの機能面での有効性が評価できたことは漢字字形情報管理システム構想の前進であると確信している。グリフウィキが将来的に漢字字形処理問題の一助となれば幸いであり、そうなるべく今後も研究を進める所存である。

（1）Windows NT 漢字処理技術協議会『Windows NT 拡張漢字処理仕様書』第二・一版、一九九八年
（2）平成十六年四月一日付法務省民一第九二八号民事局長通達
（3）上村圭介、小町祐史「ISO/IEC 10036 によるグリフ識別子登録の現状とその応用」『画電学会年次大会予稿集』、二〇〇四年
（4）Hideki Hiura, Eric Muller「Ideographic Variation Database」、Unicode Technical Standard #37、二〇〇六年
（5）小山壽久「『日本語外字センター』の提言 ネットワーク・コンテンツ時代の日本語フォントについて考える 第五回 日本語外字センターの役割と概要」『行政＆ＡＤＰ』二〇〇三年十二月号、pp.16-21
（6）WikiWikiWeb http://c2.com/cgi/wiki
（7）Wikipedia http://wikipedia.org/
（8）Jim Giles「Internet encyclopedias go head to head」『Nature』438: 900-901、イギリス、二〇〇五年
（9）Eric S. Raymond『The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary』、Oreilly & Associates、アメリカ、二〇〇一年
（10）上地宏一「漢字フォント自動生成サーバ〝影ＫＡＧＥ〟の構築 ―文字コードの枠組みを越える次世代漢字処理の提案―」『漢字文献情報処理研究 第三号』、pp.4-13、漢字文献情報処理研究会、好文出版、二〇〇三年
（11）GNU Free Documentation License http://www.gnu.org/licenses/fdl.html、第一・二版、二〇〇二年

（12）GNU General Public License http://www.gnu.org/licenses/gpl.html、第三版、二〇〇七年
（13）How does the GPL apply to fonts? : Frequently Asked Questions about the GNU GPL http://www.gnu.org/licenses/gpl-faq.html#FontException
（14）日本漢文文献目録データベース　http://www.nishogakusha-coe.net/database/
（15）ＣＨＩＳＥプロジェクト：文字に関するさまざまな知識のデータベース化 http://kanji.zinbun.kyoto-u.ac.jp/projects/chise/char-data/
16）http://glyphwiki.org/

五（四） グリフウィキ表示例



六（二） 版本表現型全文テキスト表示への適用例（枠内字が外字）