

中国古文字のデジタルテキスト化に関する諸問題

山田 崇 仁

はじめに

中国古代史の重要な研究資料は、文字資料である。文字資料は、長い間、人が書き継いで（後には印刷で）今日まで伝えられた「伝世文献」と、一旦抄写の伝承が途絶えたものが、後の時代に考古学的発掘や盗掘などの手段で現代に再発見された結果、現在の我々が認知することが可能となった「出土文字資料」の二つに区分される。

今日、中国史研究に関する史料が陸續とデジタル化されている。中でも、筆者がメインの研究領域とする先秦史の研究資料、特に正史や四書五経といった前近代中国の価値観の中心に位置づけられていた伝世文献群については、二十世紀には既にデジタルテキスト化が進められていた。その後も、伝世文献のデジタルテキスト化は進み、今日では『文淵閣四庫全書』・『四部叢刊』などの叢書や中華書局が校訂し活字化して出版した新式評点本の伝世文献類など大規模文献群がデジタルテキスト化され、データベースとして利用可能となっている。また、それ以外にも多くの文書類の目録や本文に関するデータベースが利用可能となっている（漢字文献情報処理研究会（2021）参照）。

その中で比較的立ち後れているのが、先秦期の出土文字資料を対象としたデジタル化である。無論これらの資料についても、20世紀に既に公開されている香港中文大学のデータベースなど、複数の研究成果が存在するものの、それぞれの環境や時代性故に問題を抱えており、利用にハードルが存在してきた（鈴木敦（2014））。その理由は、出土文字資料に記される文字が、漢字の古い字体、即ち「古文字」と呼ばれるデザインで描かれており、我々が通常利用する楷書体（+明朝体・ゴシック体など、そこから派生した文字デザイン）との変換において、少なからず問題が存在するためである。

本稿では、これら古文字史料のデジタルテキスト化に関する諸問題について、「古文字を読み解く過程」・「文字コードについて」・「古文字のデジタル化」の三つの項目に分けて説明する。

中国古文字を読み解く過程

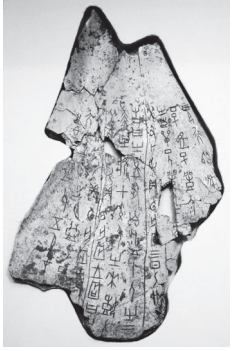
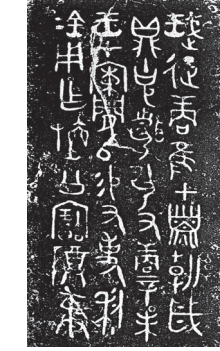
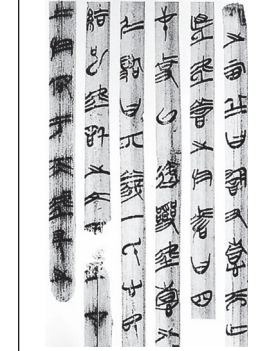

出土文字資料の特徴

中国先秦期に使用された文字（以下、古文字と称す）は、現在より約三三〇〇～二〇〇〇年以上前に使用された古い時代の中国語（上古

漢語)の書き言葉である。古文字は基本的に出土文字資料として獲得されるが、それらの多くは紙以外(骨・竹・木・絹・金属・陶器・石など)の媒体に記されており、また楷書体より数千年前のものであるため、文字の形(字体)や書きぶり(書体・書風)も楷書体とはいじりしく異なっている。それに加えて、現在では使用されない語彙(死語)も多く(死語である関係上、特に発音・意味などの部分で解釈が困難になる)、文字を読み解くのに、専門的な訓練が必要となっている。

中国学の分野では、そのような古文字を専門とする学問領域を「古文字学」と読んでいる。古文字学は、字の形や書きぶりだけでなく、発音(音韻)や意味(義)といった文字や言語に関する領域を始め、それらの古文字が使用された歴史的環境をも含む複合的な学問領域となっている。

以下に、古文字の例を挙げておく。学術用語として、殷代の骨(陸亀や牛・羊・鹿などの哺乳類)に刻まれた古文字を「甲骨文字」・青銅器に铸込まれるあるいは刻まれた古文字を「金文」と呼ぶ。戦国時代に降ると、竹や木に記した古文字が獲得されるが、この頃には地域毎に異なる書きぶりが見られるようになり、それらを踏まえて、「楚文字」・「齊文字」などと呼んだり、まとめて「戦国(古)文字」と呼んだりする。それら地域毎の書体や語彙を統一したのが統一秦期の「文字統一」政策であり、そこで制定された秦の公式書体を母体として、その後八百年以上かけて現在の楷書体に変化するのである。

			
<p>甲骨文(『甲骨文合集』137正/武丁期)</p>	<p>金文「利簋」(『殷周金文集成』4131/西周初期:成王期)</p>	<p>戦国竹簡「孔子詩論」/清華大学蔵戦国竹簡(戦国中期)</p>	<p>里耶秦簡:(秦始皇帝時代)</p>

古文字学による字形の読み解き

次に、古文字を読む（形・意味・音を解読する）ためのプロセスについて説明する。

そもそも、古文字を含む漢字と呼ばれる文字は、中華文明で 사용되는言葉のために発明された書き言葉であり、中国語（漢語）の歴史だけではなく、中華文明そのものと深く結びついている。そのため、古文字を解読するには、中華文明の歴史・論理を基盤とする必要がある²⁾。

古文字を読み解くためのプロセスは、以下の通りである。

1. 文章から一字分を切り分ける
2. 一字を構成する部品単位に分ける
3. 部品毎の字形を確定し、一字全体の形を定める
4. 音を推定する
5. 意味を推定する

1～3が隸定（字形を定める）、3～5が字積（音・意味・字形を定める）と、それぞれ呼ばれる過程となる。3が重なるのは、隸定の終着点がそのまま字積の出発点になるためである。

隸定の過程

「隸定」とは、厳密に言えば「古文字書体を隸書体」に変換することを指す。元々、唐・孔穎達『尚書正義』「隸に従つて之を定む」に

ちなんだ表現で、前漢の魯共王時代に孔子旧宅の壁を破壊した折に見つかった（おそらく先秦期の）文献を、前漢当時の隸書体に変換して整理記録し皇帝に献上した故事に由来する（『漢書』芸文志）。古い書体を現行書体に変換したからこそ「隸定」であり、現代風に言えば「楷定」とする方がよりそれらしいが（実際にそのように称する研究者もいる）、慣例に従って「隸定」と呼ぶ場合が多い。

では、実際の隸定過程を説明しよう。

まず始めに、文章から一字を区切る。漢字で書かれた前近代の文章（漢文）は、古文字に限らず基本的に句読点などの区切り記号は付与されない。隸書体以降の書体で記された漢文であれば、写本であれば印刷であれ一字が概ね仮定の正方形内に収まるように描画されるため、それで一字を区切ればよい。ところが古文字ではそう単純に処理できない場合がある。以下の金文を例にして説明しよう。



小臣守斝

（『殷周金文集成』4179、西周早期）

基本的に古文字の横幅は基本的に同一であり（縦方向の長さは字によって異なる場合がある）、前後の文字間をやや空けて配置しているため、それらの情報によって一文字毎の区切りを確定することができ。例で挙げた小臣守殷も、前後の文字とは間隔が空いており、比較的容易に一文字を判別可能である。

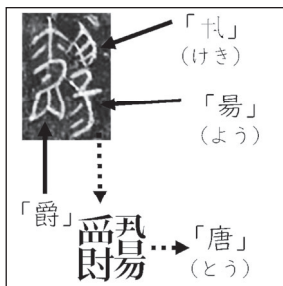
しかし、文字毎の間が極端に狭かったり前後が重なったりする例もあり、どこで一文字を区切るべきか、あるいはこの字は二文字か一文字で悩む場合もある。小臣守殷の場合、作器者_{小臣}がそれに該当する。当該字は、「小臣」二字が他の一字分のスペースに描画され、「小」の左右の点と「臣」の最上部とが水平レベルで重なっており、単純に正方形で他の字と区別することができない。ではなぜこれを二字に分けて解釈したかと言えば、他の_{小臣}を使用する銘文との比較において「小臣」が一字ずつ分けて描画されているものから本銘文も二字として区別すべきと判断したためである。

また別の例を紹介しよう。先に挙げた利簋である。この金文は、西周武王による殷討伐を記したものとして名高いが、冒頭の「武王」と解釈される字が問題となっている。当該字は_武という形だが、左側に「王」、右側に「武」という二つの部品が並ぶ構造を持つ。これを「小臣」の例のように「武」と「王」の二字に分けるか、「珅」という一字で表記するかで隸定が分かれる場合がある。武王の場合、他の金文の例では二字で書き分けるものもあるが、類似した例として、大盂鼎（『殷周金文集成』2837、西周早期）に見える彼の父文王を_文（「攸」と表記するものがある。これらの例を参照して、分割の判断をする。

文王・武王の場合、彼らが西周興隆期の君主であることを踏まえ、西周初期からこの両者の扱いが格別だったとする資料と見なし、「珅」・「攸」と表記するのも一つの考え方であり、検索の便などのために他の用例と文字表記を揃えて「武王」・「文王」とするのも一つの考え方である。いずれにせよ最終的な判断は、隸定者にゆだねられることになる。

更に、文字をどのように区切るか、あるいは文字を隸定するかが難しい例もある。小臣守殷の場合、末尾の_{子孫}が該当する。この二字は「子」と「孫」とそれぞれ隸定されるが、各字の右下に「_{子孫}」のような符号が付いているのに注目してほしい。この符号を学術用語として「重文符号」と呼ぶ。重文符号は、これが付された字を繰り返し読むことを指示するもので、「々」などの踊り字と同じ役割を担う。そのため、文字を補って表記するならば「子子孫孫」あるいは「子々孫々」となる。重文符号が付された字を解釈するときは「子子孫孫」（「子孫子孫」と読むべきとする説もある）とすればよいが、隸定する段階では「子子孫孫」あるいは「子_{子孫}」のどちらがよいかで悩むときもある。これも実際には、ケースバイケースである。そもそも論として、当該部品を重文符号と判断すべきか、あるいは字体の一部なのかについて議論が分かれる場合もある。

とにかく文字の区分けをしたら、次は字の形を定める。まずは字の形をトレースし、大まかな全体構造を理解する。次に、字をいくつかの部品に分割して個別の部品の形を検討し、古文字より後の書体（篆書体・隸書体など）ではどのような形になるかを定める。



隸定の例：晋公盤（吳鎮烽『商周青銅器銘文暨圖像集成統編』952、春秋中期）

最後に部品を組み合わせて字全体の形を確定する。ここまでは「隸定」の過程となる。現代の古文字学では、検討した部品や字全体のデザインを楷書体に変換する作業が加わり、それを含めて「隸定」と呼ぶが、上述のようにこの段階を隸定と分けて「楷定」と呼ぶ研究者もいる。

以下に、「晋公盤」の「卣」を例として隸定を試みている。まず当該字の構造を分析し、ここでは左・右上・右下の三部品で構成されると判断した。次に、各部品毎に個別検討を行い、部品毎に既知の古文字と比較し、左を「爵」・右上を「卣」・右下を「易」と定める。それらの作業の後に部品を組み合わせて全体構造を確定し（隸定）、楷書体「卣」とする（楷定）。「卣」の場合は、個別部品の分割や部品の隸定が比較的容易であるが、部品の形や全体構造が確定不明な場合は、不明字として処理する場合も多い。そのため、古文字学の分野では、不明字の解明や既知の字体の再検討（全体構造や部品の隸定の再検討・それに伴う字積の変更）が学問的検討の対象となり、特定字について複数の隸定字候補が研究者によって提示されることも珍しくない。

「隸定」作業の結果、字体が確定するが、その際いくつかの問題が生ずる。それは確定時の部品の形や配置の問題である。

「宝」を例として説明しよう。「宝」の旧字体は「寶」であり、建物「宀」に「玉」や「貝（タカラガイ）」などの威信財を貯蔵する形に從う文字である（音を表す声譜が「缶」）。ところが「寶」の古文字は、バリエーション豊かな字体を持っている。以下の例は、『新金文編』の「寶」の項目より、筆者が切り貼りして作成したものである（当該書では、「寶」だけで8ページ半を占めている）。



『新金文編』（作家出版社、2011年）の「寶」より作成

旧字体の字形に比較的似ているのは、「己侯簋」（07.37221／番号は『殷周金文集成』の巻数と整理番号）だろう。他には「是婁簋」（07.3911.1）や「伯鬲簋」（07.3774）のように「貝」が左下に入り込み「缶」

が右部分を占める「𠄎」と隸定可能な字形や、「友父𠄎」(06.3727)のように「王」「缶」の配置が入れ替わる「寶」と隸定可能な字体もある。更には「農父𠄎」(06.391)のように「王」「缶」を欠いた字体「眞」などもある。

「寶」のように数多くの異体字が存在する場合、それらをどのように隸定するのがよいか。まず、個々の字(「𠄎」「寶」「眞」)が「寶」と同一視可能かを検討する必要がある。この段階の作業は、後述する解釈とも関連する。「寶」の場合、構成部品の種類や構成要素(「ナ」「王」・「貝」・「缶」)が同一視可能か否かの基準となるが、それ以外に発音(声符)や文脈から判断する場合もある。どの字をどの字と同一視可能かについて、研究者個々にその最終的判断が委ねられている。そのため、次の解釈段階で見解の違いが生ずることになる。

そして次に、同一視可能と見なした場合、隸定・楷定する段階での作業方針を定める。この段階で、元の字毎に異なる隸定結果を尊重すべきか、あるいは同じ語を表現しているのだから現在の標準的な字体(ここでは「寶」)に統一すべきかを決定する必要がある。これは、最終的な用途により使い分けている。字体の部品配置の特徴そのものを対象とする必要性があれば、個別の隸定を行うべきだろうし、古文字資料を比較的同質なデータ群として正規化したい場合であれば「寶」と統一すべきだろう。単純な通読・積読であれば、個々の字体の隸定・楷定字を挙げ、そこに現在の標準字体をカッコ付きで表示する場合や、一律に標準字体で表示する場合に分かれる。

これら隸定作業は古文字研究の根幹を為す部分であり、個々の部品

の隸定はもとよりその集合体である字全体の隸定も、研究者の見解が分かれることが珍しくない。また、部品や字全体の解釈に見解の相違がない字であっても、隸定のデザイン(部品配置・部品のデザイン)の違いによって、最終的な楷書体字体が異なることも珍しくない。解釈に問題がない文字の隸定字体が異なる理由は、学界で統一的なルールがないためだが、それが本稿での主要なテーマである古文字資料のデジタル化に際しての問題が生ずる要因となる。

ここで個人的な楷書体置き換えの方針を述べておく。大枠として、字体を(比較的)厳密に対応する楷書体字形に置き換えるのであれば、せめて文字部品の配置は元字のそれを尊重すべきであると考えている。少なくとも一個人の研究の中で、同一字体の隸定字の部品配置などを恣意的に変更するのはさすがに問題であり、せめて同一著作の中では統一して欲しいと考えている(無論、学問的進展による隸定字体変更は除く)。

字積の過程

隸定の結果字体が定まったら字積作業を行い、発音・意味などを定める。

字積で最初に行う作業は、字を構成する部品群から「意味・領域を担当する部品(義符)」と「発音を担当する部品(声符)」を採す行為である。個別の字に関する異体字情報の事例が増えた今日、積字の作業はこのような部品の分解と各部品の役割を仮定し検討する作業の段階で、『新金文編』のような文字集に掲載される個別字体と比較する

作業が必須となる。そして、目的の文字が最終的にどれに当てはまるかについて、他の文字や関連研究を調査・分析して一応の結論を出す。

先ほどの「寶」の場合、李学勤編『字源』（天津古籍出版社、2012年）によれば、「寶」は元々（意味を担当する部品）「宀」・「玉」・「貝」のみで構成される会意字（建物に宝物・財物を蓄えた形）とされ、後に声符「缶」（上古音で帮紐幽部³）が新たに付与されたと記す（執筆…張玉金）。この見解に従えば、元々「賁」・「寔」がより古い字体だったが、後に「缶」が付与された字体「寶」が『説文解字』の見出し書体（小篆）として採用されるなど標準字体的地位となり、楷書体へと受け継がれたことになる。

上述のように「寶」には多くの異体字があるが、それらを異体字関係にあるとしたのは、義符や声符について同じあるいは類似する部品を共有している点に加え、前後の文脈から同一字として認識可能か否かについて検討した結果による。それら異体字について現行字体との結びつけが可能なものは、より古い字体や『説文解字』小篆字体を一つの目安として、部品点数が多いものを「繁（体）字形」・少ないものを「省（体）字形」と大まかに呼び分けている。その観点からすれば、「寶」は元々「繁字形」であったものが後に標準字体の座を占めたことになる。

もし字積の結果、既存の文字との関係が存在するとした場合、それがどのような理由で結びつけられるかを提示する必要がある。それを明らかにする重要な根拠が文脈上の役割であり、当該字が名詞・動詞・修飾語・被修飾語・機能語…等の何れの役割を担っているのか見定

める作業である。その上で、目的とする字の音・意味が既存の字とどう関係するか、音と意味を利用して調査・分析する作業を行う。特に古文字の場合、例えば現在は滅亡している過去の特定集団（固有名詞で表現される）のような現在死語となっているもの以外に、生きている概念であっても現在では他の語（字体）で表現される場合も珍しくない⁴。そのため、同一概念が現在他のどのような語（字体）で表現されるかを解明するのは、古文字を読み解く上で重要な作業となる。

このような字について検討する場合、漢字の多くが形声字である点を利用して、まずは音の結びつきから検討する方法が採用される。その場合、声符に該当する部品を定めた後に当該部品の音を定め、同一・類似した音（或いは声符）を持つ字から文脈上それに該当しそうな候補を挙げ、調査分析の上結論を下す。この方法を行う場合に利用するのが、この時代の漢字音（上古音）に関する研究成果である。特に昨今は出土文字資料が増加し、また既存の文献との比較可能な事例も増えた結果、同一・類似した音や意味を共有する字（通假字）の知見が蓄積され、不明字の字積に強力な道具となっている。それら通假字の情報を利用した字積を提示する場合、「それは恣意的な同定ではないか」という批判を受けないように慎重な字積作業を行う必要がある。

先に挙げた「鬪」を例にすると、まず当該字を含む部分が「晋公曰、我皇祖鬪公」であり、文脈から「鬪公」は晋国初封の君主と認められる（「皇祖」は「始祖」の意味）。『史記』などの伝世文献を調べると、当該人物は「唐叔虞」と呼ばれる人物であることがわかる。そのためこの文字は、「唐」と何らかの関係性があると推測される。そこで音

の結びつきから調べてみる。

『説文解字』の「唐」に「易」を構成要素に持つ「嗚（鳴）」が「古文（おそらく戦国時代の秦以外で使用された系統に由来する字体）」として掲示されている。「嗚」は「唐」と韻部（何れも「陽部」）を同じくする字で互いに通假関係にある。これを踏まえて「鬻」について検討すると、「鬻」は「爵」・「𠂔」・「易」の三部品で構成されるが、「嗚」と部品を共有する「易（陽部）」が声符であり、「鬻」の発音も「唐」同じか類似していたと推定され、互いに通假関係を設定可能となる。以上の検討の結果、伝世文献の「唐」と解釈して問題ないという結論となる（おそらく「鬻」は「嗚」系統の繁体と考えられる）。

これら一連の作業の結果、ようやく一字の形・音・意味について一応の見解を出すことが可能となる。古文字で書かれた文章を読むためには、その課程を繰り返す必要がある。実際の研究過程では、隸定と字積の検討が平行して進める場合が多い。より詳細な隸定・字積過程の事例は、『漢字学研究』第三号所収の馬越靖史(2015)・佐藤信弥(2015)を参照して欲しい。

以上の過程を経てもどうしても読めない字も存在し、隸定・字積の答えが出ない場合がある。古文字の中には名詞や動詞など、文法上の役割は判明するものの音も意味も不明な字や、既存の字体と明らかにかけ離れたデザインとなっており、部品の推定すら難しい字も存在する。それらについては、全く不明であるという見解を出す場合もある。また、文法上の機能や音や意味での僅かな類推など手がかりに、確言できないと断りつつ推測を提示し、他日の検討を待つ場合もある。後

世、新たな知見を得た結果、既知の字や不明字の解釈が変化する場合も珍しくない。その積み重ねが、古文字学の歩みでもある。

文字コードと漢字

文字コード

古文字のデジタル化とは、上述した隸定→釈字の過程で確定した字体をデジタル化する作業に他ならない。

現在のコンピュータでデジタル化された文字を扱う標準的方法は、文字コード(情報交換用符号化文字集合)である。⁵⁾ここでは「文字コード」並びに、「日本の文字コードの歴史」と現在主流となっている文字コード規格「Unicode (UCS)」について簡単に説明しよう。

文字コードは、「文字集合」と「符号化」という二つの要素で構成されている。

文字集合

文字を収録するための表を作成し(表のマス目の数が文字コード収録字数数の上限)に、収録したい文字を配置する作業である。各マス目(区画)は、固有の番号が付与されている。

以下に挙げた表は、日本の文字コードであるJIS X 0213 (JIS漢字コード)の1面4区の文字表である。

表：JIS X 0213 (JIS 漢字コード) 1面4区 (1-4-00～1-4-99) の文字表

	0	1	2	3	4	5	6	7	8	9
1-4-0*		ぁ	あ	い	い	う	う	え	え	お
1-4-1*	お	か	が	き	ぎ	く	ぐ	け	げ	こ
1-4-2*	ご	さ	ざ	し	じ	す	ず	せ	ぜ	そ
1-4-3*	ぞ	た	だ	ち	ぢ	っ	つ	づ	て	で
1-4-4*	と	ど	な	に	ぬ	ね	の	は	ば	ぱ
1-4-5*	ひ	び	び	ふ	ぶ	ぶ	へ	べ	ぺ	ほ
1-4-6*	ぼ	ぼ	ま	み	む	め	も	ゃ	や	ゅ
1-4-7*	ゆ	ょ	よ	ら	り	る	れ	ろ	わ	わ
1-4-8*	ゐ	ゑ	を	ん	う	か	け	が	ぎ	ぐ
1-4-9*	げ	こ								

JIS X 0213 は2つの表を持ち、それぞれが第1面・第2面と呼ばれる。個別のマス目「点」に一字が配置される。更に百「点」毎に上位単位「区」を設定する(0区～99区)。個々の「点」(〓そこに収録された文字の番号でもある)は、面・区・点の各番号を組み合わせて表現される。例えば、清音の「あ」は「1面4区2番」、同じく清音の「い」は「1面4区4番」となる。またこの表は、「は」行が「はばびひびび…」と並ぶように、いわゆる五十音図とは異なった配列順になっている。そのため、JIS X 0213で「全てのひらがな」の範囲は、拗音の「あ」(1面4区1番)から鼻濁音の「ん」(1面4区92番)までとなる。

符号化

この表のマス目の番号を、0と1の組み合わせに置き換える(デジタル化)作業が「符号化」である。文字コードでは、一般的に「文字表のマス目に付与された個別番号」をデジタル化の対象とする。

まず、何桁かのビット(バイト・オクテット)⁶⁾を単位(枠組み)とし、その枠の中で文字(実際には文字表の個別番号)を任意のビット列に置き換えるが、その際、「一つのビット列の組み合わせと文字表の一区画とが一对一で対応する」原則を守る必要がある。符号化には、番号を計算式でビット列に置き換える方法と、予めビット列に置き換えやすい16進法で表の番号を与えておき⁷⁾、それを単純にビット列に換算する方法などがある。

現在のコンピュータの主要なオペレーティングシステム(OS/基本ソフト)では、何れもこのような仕組みで文字を利用可能にしてい

る（OS側の実装）。各OSには、OS側で実装する文字コードがあり、各アプリケーションソフト（応用ソフト）は、その中から対応可能な文字コードを選択して対応する（アプリケーション側の実装）。利用者が文字コードを利用するためには、OS側の実装はもとより、アプリケーション側の実装が必要になる。

日本の文字コード

世界最初の文字コードは、アメリカで一九六三年に制定されたASCIIコードだが（安岡孝一・安岡素子（2006））、その後、日本語を始めたとする多くの言語・地域・集団毎に文字コードが文字セットの規格が制定されるようになった。

日本の文字コード規格は、財団法人日本規格協会が日本産業規格（いわゆるJIS規格）の一つとして選定・公布しており、俗にJIS漢字コードと呼ぶ⁽⁸⁾。JIS漢字コードは、一九七八年にJIS C 6226:1978が制定（78JIS）され、単一の文字表に「非漢字（ひらがな・カタカナ・記号など）」四五三字・「第1・第2水準漢字」六三四九字がそれぞれ収録された。その後、一九八三年にJIS X 0208:1987（83JIS）が制定され（総計6877字）、一九九〇・一九九七年の改正を経て収録字数は六八七九字にまで増加した。

JIS漢字コードに収録された漢字は、日常生活での利用をそれなりにまかなう数であったが、筆者のような古典中国文献を利用する人間はもとより、多くの分野から収録文字数が足りないという要望が寄せられてきた。そのため、IBM・NEC・富士通といったコンピューターメー

カーは、JIS漢字コードの空き領域（表に文字が未割り当ての部分）に独自に文字を配置した表を設定した（私的拡張）。特に、IBM/NEC拡張漢字では、コンピュータ（JIS漢字コードを実装するワープロ専用機）が一般に普及し始めた一九七〇年代末に中華人民共和国の最高実力者となっていた鄧小平の「鄧」や、日本の百貨店である高島屋の「高」などの、日本語で記される文字世界でよく使われるものが収録されていたことや、一九八〇年代～一九九〇年代にビジネス用途で日本のパーソナルコンピュータの主流期だったNECのPC-9801シリーズなどで実装された結果、これらの私的拡張字が重宝された。

JIS漢字コードの側でも従来の文字表を拡張すべく検討を重ね、一九九〇年にJIS X 0212:1990が選定された（補助漢字：五八〇一字）。補助漢字は、「日本の国語に用いられる文字」というコンセプトで文字を収録したため、筆者のような漢文を日常的に利用する利用者にとって、例えば『論語』八佾の「佾」などの文字が収録されているなど、それなりに便利だったのだが、実装面で問題があり残念ながら一般にあまり普及しなかった⁽⁹⁾。

その後、補助漢字の反省を踏まえて、二〇〇〇年にJIS X 0213:2000が制定される（JIS0213・0213JISと略称）。JIS0213は1面と2面の二つの表を持ち、97JISベースの表（第1面）の空き領域及び新たに追加した表（2面）に、それぞれ追加の文字を配置したものである。

● JIS第1・第2水準漢字はJIS X 0208:1997 準拠

● 第3（第1面の空き領域）・第4水準漢字（第2面）を新たに制定：約三六八五字

● 非漢字（記号等）の増加（第1面の空き領域）：一一八三字
JIS0213は、JIS X 0208の後継としての位置づけであり、最新のJIS漢字コードに位置づけられる。JIS0213は、二〇〇四年に大きく改訂された。その原因は、二〇〇〇年十二月に出された国語審議会答申の「表外漢字字体表（常用漢字以外の文字デザインをどのような形にするかのよりどころとなるもの）」⁽¹¹⁾によって定められた「印刷標準字体」に伴うものである。国語審議会東普では、印刷標準字体を『康熙字典』などの伝統的な字体に回帰する方針としたため、JIS漢字コードやそれを実装するためのフォントの文字デザインが変更され、従来の字形は別な面区点番号に割り当てられたものもあった（総計一六八字）。

Unicodeの選定（国際標準）

そもそも文字コードとは、コンピュータを利用して文字情報の交換を円滑にするための仕組みであり、本来ならばより多くの人が利用可能な文字集合が望ましい。ところがJIS漢字コードのような地域・言語・集団毎に異なる複数の文字コードが増加した結果、文字コード間のデータ変換や、複数文字コードの使い分け（或いは同時使用）などに問題が生じた。特に、自言語の文字コード以外で開発されたコンピュータソフトウェアを、自言語の文字コードに対応させる作業（ローカライズ・地域化）に際し、非常に手間が生ずる状況が存在し、便利なアプリケーションソフトであっても利用できない状況が存在した。⁽¹²⁾
そのため、複数文字コードで使用される文字を含んだ単一の文字コードを作成すれば、上記のような問題が解消されるのではないかと

期待し、国際標準を目指した文字コードが計画された。一つはApple、Microsoft、IBMなどの大コンピュータメーカーが中心となって結成したUnicodeコンソーシアムが制定する規格であり、もう一つは国際標準を策定する非政府機関ISO（国際標準化機構）・IEC（国際電気標準会議）の連合が制定する規格であった。その後、両者が歩み寄ってISO/IEC 10646（以下、UCS：Universal multi-octet coded Character Setと略）が一九九三年に制定された。その後は基本的に双方が歩調を合わせて改正を行っている（UCSが改定されると、それに対応した新バージョンのUnicodeが制定される⁽¹²⁾）。以下本稿では、この文字コードをUnicodeと表記する。

Unicodeでは、表1面に $256 \times 256 = 65536$ 字分が割付可能である。その表は、0面（BMP：Basic Multilingual Plane）及び1〜16面の合計17面存在する（理論上、1174112字分割付可能）。

Unicodeには過去現在にわたって使用されてきた現用文字や使われない文字が幅広く収録されているが、現状漢字が最も収録字数が多い字種となっている。元々のUnicode収録漢字は、日本・中国・台湾・韓国などの文字コードに収録される漢字を典拠としたものであった（日本の場合はJIS漢字コード）。現在では既存文字コード以外に、UCSのメンバーである国家・地域（ナショナルボディ）がそれぞれ発議元となって新規登録候補を提案し、漢字専用のワーキンググループIRG（Ideographic Rapporteur Group）／ISO/IEC JTC 1/SC 2/WG 2（国際標準化機構及び国際電気標準会議、第一合同技術委員会、第二小委員会、第二ワーキンググループ）の下部組織である）を設置して

その妥当性などについて検討作業が行われる¹³⁾。最終的には UCS の総会で候補字案を検討し収録の可否を正式決定する段取りとなる。

追加提案はナショナルボディを通じて行われる以上、国家組織関連からの提案が多いが（日本から法務省の戸籍統一文字がその一例）、一個人であってもナショナルボディに依頼して追加提案することは可能である¹⁴⁾。国以外の組織が提案母体となって新た収録された漢字集合の例として、日本の SAT 大藏経テキストデータベース研究会¹⁵⁾が『大正新修大藏経』データベース化に際して収集された Unicode 未収録字の中より提案されたものがある（CJK 統合漢字拡張 B・F・G・H の一部に追加）。

Unicode に収録された漢字は、二〇二二年十月現在、第 0 面（URO・拡張 A）・第 2 面（拡張 B～F）・第 3 面（拡張 G・H）の三つの表を使用している。

- 1993 年：CJK 統合漢字拡張（Unicode1.1 / URO）（20902 字、後に 92 字増加）
- 1999 年：CJK 統合漢字拡張 A（Unicode3.0）（6582 字、後に 10 字追加）
- 2001 年：CJK 統合漢字拡張 B（Unicode3.1）（42711 字、後に 9 字増加）
- 2009 年：CJK 統合漢字拡張 C（Unicode5.2）（4149 字、後に 4 字増加）
- 2010 年：CJK 統合漢字拡張 D（Unicode6.0）（222 字）
- 2015 年：CJK 統合漢字拡張 E（Unicode8.0）（5762 字）

- 2017 年：CJK 統合漢字拡張 F（Unicode10.0）（7473 字）
 - 2020 年：CJK 統合漢字拡張 G（Unicode13.0）（4939 字）
 - 2022 年：CJK 統合漢字拡張 H（Unicode15.0）（4192 字）
- Unicode の最新バージョンは 15.0（2022 年 6 月公開）、同バージョンに収録されている CJK 統合漢字は約 97000 字になる。またそれ以外に、互換漢字（収録元の文字コードである原規格との互換用に用意された領域）・異体字セレクタ（字体を詳細に区別する必要がある場合に利用・戸籍の異体字などに対応するため）などが利用可能。その後も、新たな追加漢字に関する議論が進められている。

古文字をデジタル化する

さて、いよいよ本稿の核心となる古文字のデジタル化作業の話題となる。

極端な話、拓本なり甲骨・青銅器・竹簡をデジタルカメラで撮影するだけ立派なデジタル化作業なのだが、本稿のようにテキスト上で古文字を利用したい場合、文字の隸定・字積の結果として確定された楷書体字形を、コンピュータで利用可能な形に落とし込む作業が必要となる¹⁶⁾。文字コード上での利用を前提とするならば、隸定・字積後の字をなるべく Unicode 収録範囲に整理したいが、上述のように Unicode には 97000 字以上の漢字が収録されており、その中には異体字関係にあるものも少なくない。どの字体を採用してどのようにデジタル化するか、さらには Unicode 未収録字はどのように処理するのかという点が問題となる。

隸定・字積とデジタル化の問題

まず、どの隸定字・字積字をデジタル化するのかという問題をとらねば。

一般に古文字の文章を資料として引用する場合、学界標準の甲骨文・青銅器銘文（金文）の大規模著録書や竹簡の個別報告書や文字集などに掲載される隸定字・字積字を入力することが基本である。そして、研究上や利用上の必要に応じて、適宜繁体形を（日本であれば）常用漢字形に置き換えたり、著録書と著者の見解が異なる場合は、その部分だけ訂正して利用することになる。¹⁷⁾

例えば金文の場合、学界標準の大規模著録書として『殷周金文集成』（1996年／全18冊、以下『集成』と略）がまず挙げられる。¹⁸⁾本書の著録銘文整理番号は学界共通で使用されるが、この著録書には金文拓本と著録情報しか掲載されない。そのため、収録金文を対象として隸定・字積（併せて「積文」と呼ぶ）を行った著作『殷周金文集成積文』（2001年／全6冊）を併せて利用する必要がある（『殷周金文集成』の修訂増補版（2007年）は、積文も併せて収録）。それ以降に獲得されたものも含む大規模著録書は、中華人民共和国・台湾からそれぞれ発行されているが、二〇一〇年代からは呉鎮烽編著『商周青銅器銘文暨圖像集成』（以下、『銘図』と略）（2012年／全35冊）が事実上標準的な地位を占めている（後に続・三編の追加シリーズも発行）。当該書では、『殷周金文集成』著録分を含めて、学界未公表の newly 発見された多数追加しており、こちらの収録番号もまた学界共通で使用されている。

金文対象にデジタルテキストを作成する場合、まずは『集成』（積文）

や『銘図』掲載の隸定・楷定されたテキストがその対象となる。¹⁹⁾ただし、『集成』や『殷周金文集成積文』の発行から既に二十年以上経っているため、積文自体の変更が行われているものも少なくない。対して新しい知見を反映する『銘図』の積文も、編者呉鎮烽独自の隸定・字積や、『殷周金文集成積文』が常用字に改めている字を、隸定字の部品配置やデザインを尊重した字体に改める場合もある。²⁰⁾

底本と文字集合の設定

古文字から隸定字・字積を定める際に部品配置の決定や異体字群から特定字を選択する行為は、個人研究であれば自身の最終的判断の下で行われる作業となるため、責任の所在がはっきりしている。ところが、学界で共有可能なデジタル資源を作成・公開したいとなった場合に、問題が発生する。何故ならば、古文字の隸定や字積は研究者によって異なる説が提示される場合があることと、新資料の獲得や学問的知見の発展により、常に新説に更新されるためであることが理由である。過去妥当であった大規模著録書や報告書の隸定字や字積が、今日では既に学界標準見解と異なる場合も少なくないため、過去の資源によってデジタルテキスト化したデータが陳腐化する可能性が発生する。更に最新の隸定・字積は、研究者間での見解が異なる場合も珍しくない。古文字学の研究営為の中で、隸定・字積作業自体が根幹部分である。そのため、文字を一意に定めることが困難な場合もある。

一研究者の立場からすれば、最新の知見を反映したデジタルテキストの提供が望ましい。しかしそれによって、入力 of 典拠となった資料

からテキスト離れすぎってしまうのも、データチェックなどの上で問題となる可能性がある。文献学の本文を扱う態度と同様に、最終的には作成者の責任において「理想の文字やテキスト」を定める必要に迫られることには変わりないが、校勘学の書物のように、何故その文字を選択したかという根拠を、どのようにデジタルデータに組み込むべきかという問題が別に生ずる⁽²¹⁾。

このように典拠となる資料間での相違以外に、デジタルテキスト(隸定字・釈字を文字コードに変換する)作業それ自体に伴う問題もある。例えば、日本語環境下で利用される状況を想定した場合、できるだけ(少なくとも釈字レベルでは)JIS漢字コード収録字の範囲に収めるのがよいだろう。その場合でも、いわゆる旧字体・新字体の字形のどちらを採用するのかという問題が出てくる。漢字学研究会が編集する学術雑誌『漢字学研究』では、「金文通解」と題して金文を読解し訳注を附す作業を毎号数本掲載しているが、この部分のテキストには旧字体を採用している⁽²²⁾。ただこれはあくまで本誌のみのルールであり、他の雑誌論文や書籍に引用される際には、新字体やはたまた簡体字・繁体字で表現されることも否定しない。日本と中国・台湾・韓国で異なる字形のどちらを採用するのか。これも、媒体の指定に応じて繁体字・簡体字の何れかを選択すればよいと考えている。

どの著録書を底本とするか、厳式(底本の自体にできるだけ忠実に表現)・寛式(ある程度常用字に置き換えて表現)の何れかを採用するか、更にとの文字集合を基本とするかなどの基本方針を設定したら、実際に入力作業を行う。著録書の選択や厳式・寛式については、作成

者の利用方針で定めればよい。学会共有資源を目指す場合でも、この辺りは作成者の方針で推進した方が作業が進めやすい。何れにせよ、底本の処理方法・使用する文字コード・文字の処理方針(厳式・寛式)などの作業方針は別途明示すべきである。また、現状採用する文字集合はUnicode一択である。Unicodeが普及した今日、ローカルな規格であるJIS漢字コードを選択する必要性は余りない。

Unicodeの符号化方式は複数存在するが、UTF-8を選んでおけばよい。作業に当たって重要な問題となるのは、Unicode文字集合の中のどの字体を優先的に採用するのかという方針である。日本語環境下で作成↓利用するのであれば、JIS漢字コードベースに第1水準↓第2水準以降の漢字↓Unicode漢字と優先順位を設定するのも一つの方針である。伝統的字体を重視するのであれば、台湾のBIG5(CNS)文字集合を最優先に使用するのもよい。

Unicodeと出土文字資料

実は古文字の(楷書化された)字形は、Unicodeにかなり収録済みである。例えば、金文の隸定・楷定字体については、『殷周金文集成積文』収録字を資料としてUCSに新規収録候補として申請し、その中のかんりの部分が主に拡張C領域以降に収録されている⁽²³⁾。拡張D以降の領域にも、例えば戦国時代の斉威王の名前「因齊(伝世文献では因齊)」の「齊(3C07C:拡張E領域)」や西周金文に見える人名「𠄎(304A6:拡張G領域)」や戦国時代の中山王の名前「𠄎(30BFC:拡張G領域)」など、金文に使用される字が追加されている。

このように、『殷周金文集成積文』が典拠となって Unicode に文字が収録されたため、以前に比べれば格段にデジタルテキストで表現（入力）できない（楷書化された）古文字は少なくなった。しかし、例えば『銘図』に新規収録された金文の新出字は当然未収録であるし、既存の字であっても研究者によって新しい隸定・字積字形が提示された結果 Unicode 未収録となる字体も存在するため、古文字を発信源とする Unicode 未収録字は常に生まれ続ける²⁴⁾。

このような Unicode 未収録字をどのように処理するか。無論、将来的には UCS への登録申請が望ましいが、隸定・字積が安定しない状態ではそれも難しい。そのため対処法的な行為となるが、現状では Unicode 外字領域へ外字として登録し文字コードの範囲内で扱う、あるいは画像を利用するなどの方法が採用されることが多い。

幸いにして Unicode 上には、外字を登録可能な私的領域が BMP 領域に加えて別に2つの表が用意されており、13万字以上を登録するだけの場所は用意されている。その表に割り付けする字の選定とフォントの実装さえ実現できれば、文字コードとしてそれら Unicode 未登録字を扱える。ただし、既に他の外字を使用している場合には、外字を切り替えて使用する必要がでてくる。それに加えて他者との情報共有を行いたい場合には、フォントや文字表を共有してもらい必要がある²⁵⁾。

近年、中国での古文字学の最新の研究成果は、Web上で公表されており、最新の隸定・積字案を提示する関係上、それらの研究成果では Unicode 未収録字が当たり前のように使用されている。そのような媒

体上で、文字コード形式で Unicode 未収録字を使用するには問題があり、画像形式で表示する場合が殆どである。外字データとして選択されるファイル形式は掲載媒体によって制約があるが、JPEG、PNGなどのビットマップ形式画像や、SVGなどのベクトルグラフィック形式画像が利用されている。論文執筆時にどうしても隸定・字積が不明な字を原稿に挿入する場合、デジカメやスキャナーで取り込んだ画像データその部分だけトリミングして加工し、原稿に挿入することも多い。画像データを利用する方法は、データベースや論文で広く利用されている。

Unicode 未収録字を文字コードの外字や画像で扱う方法には、それぞれメリット・デメリットがある。

例えば前者のメリットは、アプリケーションソフト上で文字として利用可能な点が挙げられる。ただし外字字形データを作成するためには、専門のフォント作成ソフトを利用する必要があるが、Unicodeの15・16面の私用領域について、外字作成ソフトウェアが非対応の場合もあるので、十数万字の外字を作成できない可能性もある（BMP領域に用意されている私用領域でも数千字登録可能）。

画像データは、理論上作成可能文字数の上限はない（1字11ファイルで作成）。作字には画像作成ソフトを利用する必要があり、操作方法の取得と多少の描画センスが必要となる。ただし、紙に字を描画↓それをスマートフォンなどで撮影↓画像ソフトで余計な部分をトリミングし、サイズを整えればとりあえず問題ないだろう。ビットマップ画像で作成する場合、1インチ辺りのドット数を300dpi程度には

設定しておかないと、紙面での利用には適さないものとなることに注意してほしい。また画像形式の方式上、ビットマップ画像には拡大した際に見栄えが良くないという問題が存在する。それを嫌うのであれば、ベクトルグラフィック形式のデータを採用するのがよい。こちらの方が作成ソフトの操作を身につけるのに時間を要するが、慣れれば問題なく利用できるだろう。

筆者の Unicode 未収録字対応方法

過去の文字コード未収録字処理方法の遍歴

筆者が古文字や中国古典文献のデジタル化と関わり始めてから、四半世紀以上の年月を経ている。その間、Unicode の実装・拡張などをはじめ、漢字情報処理に関する環境は激変した。そのため、筆者の文字コード未収録字に対する対応方法も、時期を経る毎に様々に変化してきた。

一九九〇年代は JIS 漢字コード (JIS X 0208) が実装の主流であったため、使用可能な漢字数は七千字に満たなかった。そのため、よく使う可能性が高い未収録字は外字を作成し、一回のみの字と判断した場合、その部分を空白で印刷して手書きで当該字を書き込んでいた。また中国語環境を実装する場合は、特殊なソフトウェアを利用したり中国語 OS を導入したハードウェアを別途導入したりするなど、高いハードルが必要であった。²⁶⁾その後、Unicode に対応した Windows NT4.0 → 同 2000 を導入した結果 (この時点でURO の 20902 字)、²⁷⁾ある程度の漢字不足を解消し、また日本語中国語混在環境を実現するこ

とができた。そのためこの時期、『史記』や『春秋左氏伝』などの伝世文献については、特定の地名や版本にのみ確認される字などの極端な例は除くと入力の不便は余り感じなかった。その一方、かつて作字した外字については、筆者がよく利用していたデータベースを利用するために、提供者が配布する外字を利用する必要があったため、それまでの外字環境とは距離を置くことになった。

丁度この時期の前後から戦国時代の出土文字資料が陸続と発見され、また金文なども併せた出土文字資料を研究材料として使用するようになると、再び外字を作字する必要性が生じてきた。外字作成の環境としては、その頃の資料整理や原稿執筆を PC 上で行っていたこともあり、Adobe 社の Indesign に実装された SING 外字ソリューションを利用していった。ところが後にこの機能が廃止されたことで、外字資産が無になる経験をする (画像データをバックアップとして別途保持していたので、画像として使い回すことは可能だった)。そのため、特定の環境のみに従った外字を利用するのは危険であると考えに至り、最近では基本的に (最終的に印刷媒体で公開する場合は) 画像で外字を作字する方針を採用している。

画像での外字制作は、拡大しても字形が荒く表示されないベクトルデータ形式の Adobe Illustrator で作成したものを利用していた。このソフトを選択した理由は、前述の SING 外字作成作業が、Adobe Illustrator で作字したデータを SING 外字に変換する作業過程であったため、その経験 (の一部) を利用したことになる。Adobe Illustrator で作成されたデータであれば印刷所で扱える可能性が高く、

ビットマップデータに比べて、データ量もそれほど増加しないという利点もある。デメリットとしては、ソフトウェアの操作を覚える必要があり、更には価格が高いために気軽に購入を進められない点がある。

2022年現在の処理方法

筆者は二〇二二年現在、Unicode 未収録字を作字する環境として、GlyphWiki (<https://glyphwiki.org/>) を利用している。

GlyphWiki は、大東文化大学の上地宏一氏が二〇〇〇年代末より開発公開する Web サイトである。その性格・位置づけは「メインページ」の解説に掲載される「グリフウィキ (GlyphWiki) は、明朝体の漢字グリフ (漢字字形) を登録・管理し、皆で自由に共有することを目的としたウィキ」である。⁽²⁶⁾ GlyphWiki は「Wiki と呼ばれるシステム (不特定多数が、共同で編集可能な Web サイトを構築するためのシステム及びその記法) を利用したシステムで、不特定多数により字体を作成・公開・共有する機能を実装・提供している。

また、GlyphWiki で作成した字形は「あらゆる改変の有無に関わらず、また商業的な利用であっても、自由に利用、複製、再配布することができま

す」⁽²⁷⁾ 「GlyphWiki: データ・記事のライセンス」より引用) というルールによって自由な利用が保証され、多くの利用者が五十万字以上の字体 (GlyphWiki では「グリフ」と呼ぶ) を登録する。GlyphWiki の利用は匿名でも可能だが、作成した字のグリフや一覧などを管理するにはユーザー登録した方がよい。

各グリフは、グリフ 1 字毎に作成される「グリフページ」で管理さ

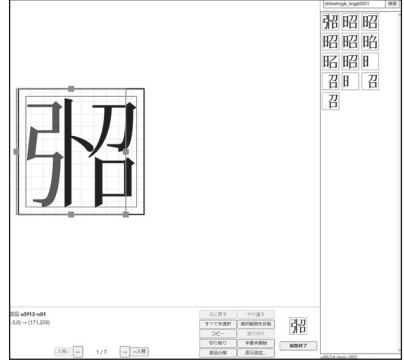
れる。グリフページには、他のグリフとの関連などの情報が掲載されるなど、文字毎に「画像 (PNG, SVG)」や「■」に割り付けられた「文字フォント」⁽³¹⁾ など、複数のデータ形式が提供されており、これらのデータ形式は汎用的なフォーマットであるため、仮に GlyphWiki が閉鎖になったとしてもデータを流用することが可能となっている。筆者は GlyphWiki に収録



グリフページの例 (https://glyphwiki.org/wiki/shrkwknjgk_knjgk0001)

者は GlyphWiki に収録されている Unicode 未収録字について、Word や Indesign での利用が可能な SVG 形式画像を主なデータ形式として利用している。

GlyphWiki の便利のところは、文字の番号を利用者が自由に設定可能かつ事実上無制限の数を登録可能な点である。無論 Unicode の文字表番号のようにシステマ的に登録不可な番号はあるが、各利用者が分かりやすい番号



kage-editor でグリフを編集している画面

を使用可能な点は便利だろう。⁽³²⁾ また上図のように、一論文でのみしか使用されない可能性がある隸定字・釈字であっても気軽にグリフを登録できるのは、筆者のような古文字を多数扱う立場からすると大変にありがた。⁽³³⁾

まずグリフ毎に割り当てられる Web ページを作成し、字形を描画すればよい。作成には GlyphWiki が付属の kage-editor を利用するのが便利である。kage-editor では、既存の文字・部品（漢字を構成する各パーツ）を参照して読み込ませたり、手書きで一から筆画を描画することができる。初めのうちは使用に少々戸惑うだろうが、慣れれば一字十分も要さずに作成できる（むしろ、作成したいグリフが他のユーザーによって既に作成済みか否かをチェックする方に時間を要する）。このようにして作成された大量の GlyphWiki データは、字形を個人が利用する以外に他の Web サービスへの字形情報提供などでも利用されている。⁽³⁴⁾

例えば Unicode では、各字形情報ページに掲載されている他の字形情報源リンク情報に GlyphWiki へのリンクを掲載して利用者の便宜を図っている。また、京都大学人文科学研究所・守岡知彦氏の作成・運

営する漢字検索サービス CHISE IDS FIND (<https://www.chise.org/ids-find>) では、Unicode 収録字について、当該字形のフォントをインストールしていない利用者向けの字形例示用データとして GlyphWiki データを利用している。更に、花園明朝フォント (<http://fonts.jp/hanazono/>) は、GlyphWiki 収録字形から Unicode 収録文字を集録し、フォントとしてまとめたものである。また、台湾の陳信良氏が作成・運営する漢字・文献の検索サービス「引得市 index (<https://www.meibag.com/index/>)」では、GlyphWiki の字を文字表記に利用する一方、出土文字資料データの検索に際し、自ら多数の外字を作成している。無論それらの GlyphWiki で作成された（隸定・釈字後の）出土文字資料外字は、他のユーザーにも公開・

引得市

NO.	書名	詞頭/字頭	巻別	分巻頁碼	字號(詞序)	合訂頁碼	漢語大詞典
1	大漢和辭典	臙	09	0380	30003	09734_14	link
2	大漢和辭典	臙壯	09	0380	30003-0001	09734	link
3	大漢和辭典	臙子	09	0380	30003-0002	09734	link
4	大漢和辭典	臙肥	09	0380	30003-0003	09734	link
5	大漢和辭典	臙滿	09	0380	30003-0004	09734	link

第 1 頁 / 共 1 頁 (總計 5 筆)

PDRC 造形藝術與資訊處理技術研究中心
The Plastic Arts & Data Processing Research Center

●説文解字本文綜合検索

【臙】 30003 ㄅㄨˋ (集韻) 惡嬌切

肥 そとやう (集韻) 臙 脂肥兒。(六書故) 臙 肥盛也。

【臙壯】 そとやう 肥えてたつて、強健。「福惡全書」潘任部、查交代「臙壯齒小、馳快健者。」

【臙子】 そとやう 肥枝をこふ。婁子(9-6397-1)の【臙肥】より。肥える。「福惡全書」郵政部、總論「專擇臙肥健馬、鞭箠。」

【臙滿】 そとやう ふとる。肥える。「王僧、賣牛詩」骨隱隱新滿、踏強齒尚差。

引得市 index の『大漢和辭典』検索で「@臙 (U+268E)」を検索した結果と『大漢和辭典』の該当部分

共有されており、筆者も日頃から大変重宝している。更に、引得市indexでは、それ等の外字をUnicodeの外字領域に独自基準で配置した外字ファイルを提供している（出土文字資料外字など、一部の有料サービスに登録後に利用可能）。

まとめ

以上、シンポジウムで話した内容を基礎に、時間の関係で省略した部分も含めて述べてきた。

二〇二二年の今日、Unicodeの登場と数度の拡張によって「漢字不足」問題は大幅解消された。しかし出土文字資料は、現在では使用されない語（死語）や字体で描画されている関係上、その隸定・字積自体が学術的研究対象であり、そのため文字コードで表現・対応しきれない字種であった。そのため、そのデジタル化に際しては、筆者も含めてさまざまな試みがされてきた（山田崇仁（2018））。

本文でも述べたように、現在ではGlyphWikiを利用して外字を作字し、それを共有するのがよりベターな形ではないかと考えている。GlyphWikiの各グリフはそれぞれ固有の文字列によって他のグリフと区別されており、その文字列を文字を表現する固有番号として利用・共有することで、画像やフォントが使用できない環境でもUnicode未収録字の共有が一応可能である。一定の前処理は必要となるだろうが、データベース作成やテキストデータ分析への利活用も可能だろう。

古文字を扱う以上、新出資料の獲得や新たな隸定や釈字によって新しい字体が生まれ続けるのは宿命として受け止めるしかない。それを

デジタルな形で扱う場合、どのような形がよりベターなのか。現在の所、GlyphWikiの文字番号とSVG・PNG画像で処理するのが比較的ベターだと考えている。また、将来的には新たな解決方法が提供されるかもしれないが、SVGやPNG形式の画像はある程度の未来まで使われ続けるだろうし、文字で代替表記できる形式を別途用意しておくことで、処理方法が異なってもコンバートが容易となる。そのため、将来の持続可能性を考慮に入れても、それなりに良い方針と考えている。

注

(1) 甲骨文字の中に、竹簡（あるいは木簡）を紐で束ねた形に由来する「冊」字が既に使用されており、殷代より簡に文字を書く習慣が存在した可能性が考えられているが、考古学的発掘によって得られた確実な資料は今のところ存在しない。

(2) 無論このような考え方は、それぞれの古文字がどのような発想で描画されたかという、いわゆる字源を考える上でも従うべきものである。昨今、文字（漢字）の字源を解説する書籍やWebサイトが複数存在するが、それらの中に当該字を楷書体の字体の各部品に分け、各部品を日本語（の中で与えられた）の意味を組み合わせて（或いは恣意的に）解釈する場合があります。このような行為は、漢字の歴史を全く踏まえていない行為であり、古文字の字形や上古音を踏まえて検討を行う古文字学の観点からは全く賛成できない。無論、本文でも述べたように、古文字学の解釈において、形・音・意味のいずれにせよどこかで踏み込む必要があるが、それは当然、中華文明と古文字学の文脈の中で行われるべき作業であり、それらを無視したものには学問的な価値は全く存在しないも同然である。

(3) 「上古音」は、後漢以前の古い中国語（上古漢語）の音韻体系を指す語。●「紐▲部」は、上古音の音韻を表現する定型句で、声母（頭子音）を「●紐（母）」、韻母（介音・主母音・韻尾）を「▲部」と呼ぶ（●・▲は、各声母・韻母のそれぞれ代表字で表現される）。上古音についての解説は、漢字文献情報処理研究会（2021）の16章「音韻を調べる」（野原将揮担当）。

- 村上幸造 (2018) を参照されたし。
- (4) 例えば、現在では「数の単位(千の十倍)」や「非常に多い」という意味で使用する「万(萬)」だが、元々はサソリを上から見た形を象った象形字であり、音の類似から上記の語を意味する表現として借りて使用されてきた。その代わりに、サソリそれ自体を表現する字として「蠍」が別途創出されたのである。
- (5) 文字コードについての詳細な解説は、深沢千尋 (2011) を参照。また、日本や東アジアの文字コードについては、少し古い内容だが三上喜貴 (2002) や安岡孝一・安岡素子 (2006) が有用。
- (6) 「ビット (bit)」とは、2進法で動く現在のコンピュータが処理可能な最小の情報単位(2択)を指す(2進数の1桁≦0と1の二種類を処理可能。文字のように種類の多い情報量を処理するために、bitを複数の桁にまとめて1単位とし(8bit ≦ 1octet, 1byte))、その中の延べ数の範囲(1octet ≦ 256bit, 2octet ≦ 65536bit) で文字が区別可能な仕組みを採用している。
- (7) 16進法は1桁を0〜9、A〜Fの合計16種類の文字で表記する方法。2進法との換算が10進法に比べて簡単であり、情報処理関連の表記法として利用される機会が多い。
- (8) 日本と同様に東アジア諸国でも、例えば中華人民共和国ではGB2313、GBK、台湾ではBIG5、CNS、大韓民国ではKS X 1001などの独自文字コードが制定されている。
- (9) 補助漢字の文字表に収録された字は、日本がUnicodeに収録を申請して可決された結果、現在ではUnicodeに収録され、利用が可能となっている。
- (10) 表外漢字字体表 解説(文部科学省) / 2022年5月5日閲覧、以下同じ。
https://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/kijun/sankohyogai/index.html
- (11) 中国語の場合でも、漢文文献データベースが1980年代から存在したが、繁体字中国語専用ソフトウェアであったため、繁体字中国語で動作するPC・OSを用意する必要があった。Yanada Takahito (2022) 参照。
- (12) UCSは日本でもJIS X 0221として規格化されているが、実装上Unicodeを使用したOSやアプリケーションソフトが圧倒的であり、また上述のように、現在では双方が歩調を合わせて改正を行っている関係上、ほぼ同一視してよい。
- (13) ISO/IEC JTC1/SC2/WG2/IRG
 Ideographic Research Group (<https://appsvn.cse.cuhk.edu.hk/~irg/>)
- (14) 日本の変体仮名がUnicodeに登録された経緯のように、日本以外のナショナルボディを通じて提案も珍しくない。また、エジプト象形文字・くさび形文字・突厥文字のように、現在既に一般的な使用者がおらず、専門家のみしか使用しない文字であっても収録対象になっている。
- (15) <https://idtk1.u-tokyo.ac.jp/SAT/>
- (16) 楷書体ではなく、古文字デザインそのままにデジタル化(文字コード化)をしたい需要もあり、UCSでも第3面にスペースを用意して中国古文字の文字コード化を試みたが、現在に至るまで作業は進んでいない。最初に試みられた甲骨文字の場合、基準となる字体デザインそのものが同時代資料として確認できないという資料上の制約などの問題があって作業が難航し、二〇一二年にプロジェクトチームが解散している。鈴木敦 (2014) 参照。個人の責任で正書法を定義できる環境であれば、落合淳思 (2006) のような試みもある。
- (17) 学界では、古文字やそれによって書かれた資料を提示する場合、「隸定字(字積)」のようなスタイルで元の字体を尊重しつつ、引用者がどのように解釈したかを分かりやすく明示する工夫をしている(上述の「鬻」ならば「鬻(唐)」と表記する)。
- (18) 甲骨文字の場合は、『甲骨文合集』(1983年/全13冊)が標準的大規模著録書であり、整理番号が学界共通のものとなっている。また、『合集』の積文については『甲骨文合集積文』(1999年、全4冊)が発行されている。更に『甲骨文合集補編』(1999年/全7冊)など、それ以外の大規模著録書に収録された甲骨文の拓本・積文を最新の知見を踏まえて収録した『甲骨文校釈総集』(2006年、全20冊)なども発行されている。
- (19) 台湾の中央研究院歴史語言研究所が作成・公開する「先秦甲骨金文簡牘詞彙資料庫」(https://inscription.asdc.sinica.edu.tw/c_index.php)では、金文の典拠として『集成』と『新收殷周青銅器銘文暨器影彙編』(2006年、全3冊)が挙げられている。ただし、このデータベースで検索可能な積文の中には、最新の知見によって底本と異なる字積に改められたものがある。
- (20) 例えば、『集成』1742の積文「亞幸獲」について、『銘図』では「亞幸(幸)憂」¹⁹⁾と積文を改めている(『銘図』では惑説として「憂」も挙げる)。
- (21) この問題は、文献を扱う中国学諸分野でも変わらない。この分野では、文献を地上で書き継ぎ(または印刷して)現在に伝わったものを主な研究材料とするが、文献は長年の複写作業によって少なからず異同が存在するため、それを校訂する作業が必要となる。最終的に、研究者個人の責任において「理想の本文」を作成する必要がある。
- (22) これは、伝世文献の表記であっても同様である。実際には、最終的な掲載媒体(学術雑誌などの専門誌/書・一般誌/書)のルールの範囲内で

表記が整理されるのが一般的。

- (23) JIS X 0221 : 2014 (ISO/IEC 10646 : 2012) 34頁参照 (<http://kikaku.rim.com/x0/X0221-2014-01.html>)。
- (24) 未収録の理由は、そもそも UCS に未申請である、あるいは申請済みであったとしても UCS での審査に時間を要している可能性がある。そもそも申請に際しては、典拠となる資料を提示する必要がある。古文字の分野では、研究者間で見解が異なる隸定・釈字もあるため、隸定・字釈が常に定まらない場合もあり、文字コードのような共有可能な文字集合と相性が悪い（研究者個々が文字集合が異なる場合がある）。
- (25) かつてエーアイネットが開発・販売していた今昔文字鏡は、JIS 漢字コードの文字表に独自に『大漢和辞典』番号順＋独自文字集合を配置した特殊な文字集合を作成・提供していた。その中には、白川静『金文通釈』由来の JIS 漢字コード未収録字が収録されていた（その他の文字も含めて、いくつかの文字について UCS 拡張漢字の典拠の一つとなっている）。現在でも古文字の楷書体字形を印刷媒体などで利用する際、採用される場合がある。
- (26) この辺りは漢字文献情報処理研究会『電腦中国字』（好文出版、1999年）を参照されたし。
- (27) <https://glyphwiki.org/wiki/GlyphWiki:メインページ>
- (28) 「GlyphWiki:グリフウィキについて」(<https://glyphwiki.org/wiki/GlyphWiki:グリフウィキについて>)によれば、上地氏の研究の一環として運用しているが、できるだけ長く運営可能なことを希望し、多くの外字情報を集約可能な手段の一つになることを目指している、とのことである。
- (29) <https://glyphwiki.org/wiki/GlyphWiki:データ・記事のライセンス>
- (30) 「GlyphWiki:データ・記事のライセンス」によれば、「グリフウィキに投稿した記事はグリフウィキ（運用者である利用者:kamichi）に対して著作権の譲渡を行ったことになり、その記事がいかなる形態で改変・利用されることも投稿者は許諾し、以後著作者人格権を主張・行使しないことを了承するものとみなします」という文の下に、自由な利用が保証されるとする。
- (31) グリフデザインを **■** (U+3013) に割り付けた 1 文字で構成されたフォントを作成する機能。当該部分を MS 明朝・ヒラギノ明朝などの他フォントに変更すれば **■** で表示される。仕組みとしては、今昔文字鏡が **SHI** JIS 上の第一・第二水準部分で採用したものを、**■** 1 文字に特化している形に近い、1 文字フォントをインストールして Microsoft Word で異体字の【西】を表示させる際の手順」(<https://note100yen.com/en-131008.html>) も併

せて参照されたし。

- (32) それ故に、システムのページの名称に複数ルールが混在する弊害もある。これについては、GlyphWiki で提供される「エイリアス」と呼ばれる仕組みを利用して、複数利用者のグリフデータを借りて別の統一されたルールに沿ったグリフページ群（文字表）を作成可能。

- (33) 論文では、複数の研究者の隸定・字釈を引用しつつ自らのそれを提示するような場合も多く、それらの字は大抵一度しか使用されない場合も珍しくない。そのため嘗ての活字・写植印刷の時代では、外字の作成コスト（外字を特別に鋳造・作字する）も鑑みて、気軽に使用しがたい状況にあった。

- (34) このように非常に便利な GlyphWiki だが、問題はこのサービスが上地宏一氏という個人の物的・人的貢献によって根幹が維持されている Web サービスという点である。本文で後述した CHSE IDS FIND・Unicode・花園明朝・引得市 index での GlyphWiki 由来のデータ利用など、既に文字コード界限において GlyphWiki は重要なインフラの一部となっている。またこれら二次利用の CHSE IDS FIND や引得市 index も、文字を扱う上で重要な存在だが、こちらも個人プロジェクトである。これらデジタル人文学を支える重要プロジェクトは、個人の運営であることは何れ大きな問題となる。長期持続的な利用を担保するために、なんらかの公有の維持管理の仕組みが提供されることが望ましいのではないかと。

参考文献

- 落合淳思「日本語用文字コードに対応した甲骨文字フォント製作案（附・甲骨文字の部首整理）」『立命館東洋史学』29、2006年。
- 漢字文献情報処理研究会編『デジタル時代の中国学リファレンスマニュアル』（好文出版、2021年）。
- 佐藤信弥「金文学のツール」裘錫圭「西周銅器銘文中的“履”を例として」『漢字学研究』3、2015年。
- 鈴木敦「甲骨文字研究の成果蓄積とデジタル化技術」（飯島武次編『中華文明の考古学』、同成社、2014年）。
- 深沢千尋「文字コード「超」研究」改訂第2版（ラトルズ、2011年）。
- 馬越靖史「金文の「易」字」『漢字学研究』3、2015年。
- 三上喜貴「文字符号の歴史—アジア編」（共立出版、2002年）。
- 村上幸造「通假字を見るための上古音概説」（『漢字学研究』6、2018年）。
- 安岡孝・安岡素子「文字符号の歴史—欧米と日本編」（共立出版、2006年）。
- 山田崇仁「關於建構殷周金文標準字元編碼的基礎研究」（『世界漢字学会 第

六届年会予稿集』(2018年)。

Yamada Takahito “The History of the IT Environment of Sinology in Japan”
(“*The International Journal of Chinese Character Studies*”, 4-2, 2022)。

胡佳佳「古籍数字化中的汉字信息处理」(『民俗典籍文字研究』(北京師範
大学) 8, 2011年)

李守奎「曹沫之陣」之隸定与古文字隸定方法初探」(『漢字研究』 1, 学苑
出版社, 2005年) 後『漢字学論稿』(人民美術出版社, 2016年) 所収。

謝辞

本稿は、JSPS 科研費 22K12738 の助成を受けたものである。

(立命館大學白川静記念東洋文字文化研究所客員研究員)