Vol.2013-CH-97 No.4
2013/1/25

UCS 符号化提案におけるデジタルツールの活用
ー大正新脩大藏經外字の符号化提案にあたって

永崎研宣[†]　清水元広[†]　下田正弘[††]

コンピュータ上で外字を扱うことは，人文学研究，とりわけ古典を扱う場合には避けることが難しい事柄である．やみくもに外字を増やすことは望ましいことではないが，字形の違いをなかったことにしてしまうことには潜在的な問題がある．したがって，包摂を前提とする UCS 符号化文字集合をそのまま全面的に導入することは困難である．SAT 大蔵経テキストデータベース研究会では，外字の扱いを巡って検討を重ねてきており，近年普及しつつある IVS を利用することで UCS 符号化文字集合の適切な利用方法を模索することを開始すると同時に，約 3000 字の漢字に関して IRG に対して符号化提案を行った．本稿では，そこに至る検討過程について報告するとともに，外字を UCS 符号化提案した際の具体的なデジタル技術の活用の仕方についても紹介する．

Application of Digital Technologies for a Proposal of UCS Encoding of Ideographs Included in the Taishō Shinshū Daizōkyō

Kiyonori Nagasaki[†]　Motohiro Shimizu[†]　Masahiro Shimoda[††]

As a result of the process of encoding all texts in the primary canonical collection of Buddhist scriptures composed in classical Chinese, called the “Taishō Shinshū Daizōkyō,” (collection of Buddhist canonical texts compiled during the Taisho era) the SAT Daizōkyō Text Database Committee began to address UCS encoding of unencoded ideographs contained in this collection. This paper describes the usage of digital technologies for this work and discusses some technical issues concerning unencoded characters in the humanities.

1. はじめに：デジタル化とデジタル翻刻

テキスト資料をデジタル化するには，現在のところ，写真を撮って画像として保存する方法と，何らかの文字コードに準拠してそのコードポイントを記述して保存する方法とがある．双方ともに長所短所があり，媒体の性質や予算規模，プロジェクトの方針等によってどちらか，あるいは両方が選択されたりすることになる．高精細な画像を撮影・保存・公開するためのコストが著しく下がってきている現在，前者のメリットは高まりつつあり，前者に取り組むプロジェクトはますます増加しつつある．また，前者を行って公開した後に後者に取り組む，さらには，後者を行うべく広く呼びかけるといったプロジェクトも進められるようになっており，すでにその手法で大きな成果を挙げているプロジェクトもある[a][1]．とはいえ，外字の問題は主に後者において大きく影響することであるから，ここでは後者を「デジタル翻刻」と呼んで特に採り上げて検討したい．

デジタル翻刻では，テキスト資料における文字をコンピュータ上で効率的に扱えるようにするために，なんらかの文字エンコーディングに準拠する形で記述することになる．かつては JIS X 0208 に基づく文字エンコーディングが用いられており，利用可能文字種の制約が大きかったが，近年では Unicode に基づくものに切り替わりつつあり，その Unicode の収録文字数もますます増加しつつある．漢字に関して挙げるなら，すでに CJK Extension D までで約 75000 字，これに加えて IVS (Ideographic variation sequence) が約 8000 字使えるようになっており，さらに，現在準備中の Extension E が約 5700 字，新たに登録されることになっている．現在のデジタル翻刻では，外字を使わずとも国際標準として符号化されている漢字だけで 80000 種類以上の字形の使い分けを前提とすることができるのである．しかし，それでも，本稿で主題とする『大正新脩大藏經』[2]（以下，大正蔵）をデジタル翻刻するには十分ではない．まずはその事情についてみてみよう．

2. 大正蔵のデジタル翻刻

2.1 大正蔵の文字の来歴[b]

大正蔵は，テキスト部分 85 巻，図像部 12 巻から成る仏典のシリーズである．インドで記述されたものを漢文に訳したもの，中国で書かれたもの，日本で書かれたもので構成されており，テキスト部分に含まれる文字数は約 1 億 5000 字となっている．この大正蔵の文字は，高麗版大蔵経

[†]一般財団法人人文情報学研究所
[††]東京大学大学院人文社会系研究科

a) http://www.transcribe-bentham.da.ulcc.ac.uk/

b) 大正蔵の刊行に至る事情については筆者が別稿にてまとめたものがあるので参照されたい．[3]

再彫本[c]（以下，高麗藏）と呼ばれるものを底本にし，宋代，元代，明代においてそれぞれ開版された木版大蔵経と校合しつつ，さらに，正倉院聖語藏をはじめとするいくつかの古写経をも比較対校したものが基本となっている．ただし，高麗藏及び宋・元・明の木版本の校合に関しては，明治期に活版印刷の和装本として刊行された大日本校訂大蔵経[4]（縮刷藏と通称されるものであり，以下，縮刷藏と呼ぶ）において行われたものを基本的に踏襲した形となっている．高麗藏はそもそも木版であることから活字に比べると字形の統合への必然性はさほど高くなく，どちらかというと異体字を多く含んだテキストであり，この高麗藏の異体字を収集した字典[5]では約 3 万字が収録されているほどである．このような大蔵経における字形の揺れは，活版印刷の和装本として出版された縮刷藏の段階である程度吸収された模様である．続く大正藏の校合作業においては，これらの木版大蔵経に加えて，日本に伝わる古写経を参照しつつ，主に日本撰述部として，高麗藏には未収録の仏典をも多く収録した．そしてその際，異体字表を新たに作成してさらなる漢字の統合が行われたのである[6]．大正藏は，西洋文献学の手法を援用しつつ大規模な校合を行い，しかも洋装本として刊行されたということで，世界中に広まり，その後現在に至るまで，仏教研究の基盤として常に参照されるものとなっている．ただし，3 段組で一冊 1000 頁近い漢文の書籍を毎月一冊くらいのペースで刊行するという仕事であったため，完璧を期すことができたわけではなく，早い段階から様々な問題点も指摘されてきている[7]．たとえば，よく知られている比較的大きな問題としては，『大乗中観釈論』の後半部分が完全に欠落しているというミスがある[8]．もちろん，研究資料としては，問題点が周知されていればそれを踏まえて次に進めば良いのであって，そのような問題点自体は研究分野全体として見渡した場合には十分に補完可能なものであり，実際にそのようにして進められてきている．

2.2 活字化の際の意図をどう扱うべきか

SAT 大蔵経テキストデータベース研究会[d]（代表：下田正弘．以下，SAT）によって 1994 年に開始された大正藏のデジタル翻刻は，大正藏の編纂過程に戻っての再検討を要求する新たな問題を提起することとなった．SAT のデジタル翻刻活動において発見された外字は約 1 万字程度であり，具体的には今後も一層の精査が必要だが，そのうち，4000 字超の外字は，他の文字と，具体的には，既存の UCS 符号化文字や大正藏に登場する他の文字と，ごく微細な字形の違いがあるのみであり，UCS の包摂規準では同じ文字とみなされざるを得ないものとなっている[e]．しかし，UCS において同じ文字とみなされ得る微細な違いを含む文字は，大正藏の編纂過程においては，果たして，同じ文字だという認識で活字化されたのだろうか．万が一，敢えて違う文字だという意味を込めて微細な違いを表現したのだとしたら，あるいは，当時の編纂者達が属するコミュニティではその微細な違いに意味があることが当然の前提となっていたのだとしたら，それらを統合してしまってよいのだろうか．このような疑問は，人文学において漢字を扱う研究者には広く共有され得る問題であり，特に，大正藏のようにすでに広く使われているテキストの場合には看過できない問題となる．というのは，すでに，それらの微細な違いを含む異体字を異なる文字として扱っている研究者もいるかもしれないからであり，あるいは，今後の専門家による研究から，その微細な違いに意味があるかどうか，さらに，場合によっては，その違いに大きな意味があることが判明するかもしれないからである[f]．Unicode の符号化文字数の増加と Unicode 対応アプリケーションの普及により徐々に使われなくなりつつあるとは言え，文字鏡フォント[g]や GT 書体[h]といった多漢字フォントが登場した背景，そして未だに局所的に用いられている理由の一つには，そのようにして，微細な違いを残しておくことで，もはやコンピュータ上でしかテキストを読まなくなってしまうかもしれない今後の研究者に，その問題があることを伝え，判断を託していこうという企図があったのだろうと思われる．そしてまた，そうした微細な違いを残しておくことは，すでに広く行われているテキストマイニング等の，コンピュータを用いてテキストから何か新しいものを見いだそうとする研究に対しても，新たな発見の余地を残しておくことができるという点で，大きく期待されるところである．

2.3 UCS 符号化提案以前の SAT での異体字対応

SAT では，この微細な違いに関して，可能な限り字形に即した形で UCS 符号化文字から同じ字形を選択するようにしつつ，そこに見つからなかった字形に関しては独自の文字番号をつけるという形で記述し，表示に当たっては GT 書体，及び GT 書体にないものは GT 書体をベースとした独自の文字画像を作成することで対応してきた．

ただ，このようにしてデジタル翻刻において異体字をなるべく残そうとすると，かつては，検索用途における不便さが指摘され批判されることがあった．たしかに，異体字を多く含むテキストを単純に検索しようとする場合にはその通りである．しかし，情報技術の進歩や情報資源のオー

c) http://kb.sutra.re.kr/ritk_eng/index.do
d) http://21dzk.l.u-tokyo.ac.jp/SAT/
e) 活字資料のデジタル翻刻に際して独自の包摂基準を設けて取り組んでいる研究もある．[9]

f) なお，このような問題は，仄聞する限りでは，コンピュータを執筆道具として用いるようになってからの一次資料からの翻刻においてはそれほど大きな問題とはならず，むしろ翻刻を行う研究者の方で，作成時・作成後の利便性を考慮して，既存の文字コードにおける文字にあてはめるという意識で翻刻に取り組む傾向があり，外字や異体字はそれほど問題にならないということである．これについては別途，機会があれば調査してみたいと考えている．
g) http://www.mojikyo.org/
h) http://www.l.u-tokyo.ac.jp/GT/

プン化により，異体字を検索の際に横断検索するような仕組みが比較的簡単に利用できるようになってきている．また，SAT の Web サービスがそうであるように，異体字の同時検索を行ってくれるサービスもでてきている．これには異体字シソーラスのデータが様々な形で提供されるようになってきたという背景があり，個人でもこうしたデータを入手した上で，テキストをいったんよく使われる文字に変換してから検索するか，検索ソフトウェアの方であいまい検索をできるような工夫をしておけばよいのであって，利便性という観点では，もはや異体字の微細な違いを残すことは障害にはならなくなってきているのである．[i]

なお，異体字を同時検索する場合には．検索にかなりの計算コストがかかることになってしまうが，SAT のような Web サービスにおいてはサーバ側にある程度大きな計算能力を持ったコンピュータを用意することで実用上の問題は回避可能である．実際の検索手法としては，あらかじめ異体字を正規化した検索用データを用意しておいて検索キーワードも正規化した上で検索する方法と，異体字データベースを使って検索キーワードを利用可能性のあるパターンに展開して異体字を含むテキストをそのまま検索するという方法がある．検索時点でのコストは明らかに前者の方が低いが，後者は検索データを用意するコストを低く抑えることができる．特に，異体字シソーラスが安定しないうちは後者の方が運用しやすいだろう．

2.4 独自の外字システムの問題点

しかしながら，やはり独自の文字番号と文字画像の組み合わせを用いるというのは実用面で取り扱い上の障害が多かったことから，次のステップとしては，文字画像をフォント化して見栄えをよくするとともに，何らかの形で異体字の使い勝手を高めることが検討されてきていた．とりあえずは，コラボレーションでフォントを作成可能な Web アプリケーション，GlyphWiki[j][10]を用いて UCS 外字フォントの作成を進めつつあった．とは言え，フォントを作った場合にも，そのままではコードポイントの扱いに問題があった．

Unicode においては PUA (Private Use Area)と呼ばれるコードポイント領域で好きな字形を自由に扱うことが許容されている．このため，敢えて UCS として符号化せずに PUA を利用して独自のフォントを作成して流通させるということも場合によっては考えられる．しかし，特に大正蔵の場合，中国語圏でも広く使われることになるが，中国語圏では PUA の利用が比較的盛んであり，したがって，SAT が PUA を用いて外字をコード化してそれに対応するフォントも作成した上で大正蔵のテキストを提供した場合，ユーザがそれと気づかずに別のフォントを使ってしまってまったく異なる字形が表示されてしまうということが十分に考えられる．そうした事態を避けるためには，やはり，UCS 符号化文字として符号化しておくことが必要となる．それが実現できたなら，流通可能性だけでなく持続可能性という面でもメリットが大きく，また，大正蔵は仏教学における共通のプラットフォームとして国際的に広く流通していることからも，UCS として符号化した場合の波及効果もきわめて大きいことが期待される．しかし一方で，可能な限り異体字を残しておきたいという SAT の方針は，UCS の包摂基準とは相容れないものであるという悩ましい状況であった．これを解決できる可能性が見えてきたのが，IVS (Ideographic Variation Sequence)[11]の実用化と普及である．このことは，SAT のおかれたそのような状況を前に進める大きな原動力となった

2.5 IVS の活用による異体字記述の可能性

IVS は本来，同じ文字に関する異体字字形を UCS 符号化文字とは異なる階層で扱えるようにするための枠組みであると考えられる．ここには，実際に漢字の符号化に取り組んできた IRG (Ideographic Rapporteur Group)会議[k]が主に採ってきた，字書に依拠して符号化するという手法における，漢字の同系・非同系という議論と深く関係があるものと思われる．しかしながら，SAT が対象とする文字の場合には，一部に音義書等も含まれるとは言え，基本的には，文字の定義等について書いたものではなく，仏典であり，いわば，用例しかないという状況である．個々の用例において用いられている外字が同系であるか非同系であるか，さらには，外字と，それに似た UCS 符号化文字が同系か非同系か，といったことは，用例の解釈に依存することになるため，その仏典の専門家がある程度時間をかけて検討しなければ適切な判断は難しい．また，後に研究が進展することで別の解釈が出てきて，同系とされたものが非同系になってしまう可能性，あるいはその逆もあり得る．このような状況では，IVS と UCS 符号化文字との関係は，大正蔵の場合には，一般的な UCS 符号化文字とは少し異なるアプローチをとらざるを得ないだろう．すなわち，確実に非同系であることが明らかな文字，字形の異なる文字に関しては UCS 符号化文字として提案するが，同系か非同系かあいまいな文字に関しては，同系の異体字と一緒に IVS に登録してしまうという仕方である．このような仕方であっても，IVS として字形の違いが標準に採り入れられるのであれば，ルールとしての持続可能性は十分に担保され，また，IVS が今後広まっていくとともに流通可能性も高まっていくだろう．IVS の当初の意図とは異なる使用法となる可能性はあるにせよ，字書にない文字を符号化する際の一つの考え

i) ただし，文字鏡フォントに関しては，文字番号と Unicode 等の別のコード番号の対照表を作成公表することを禁じていたことがあり，この場合にはコンピュータ上での様々な利用に問題を生ずる可能性がある．なお，現在の文字鏡フォント及び関連情報の利用許諾条件については把握できていないのでご教示いただければ幸いである．

j) http://glyphwiki.org/

k) http://appsrv.cse.cuhk.edu.hk/~irg/

方として，字形の相違が微細であり同系・非同系が曖昧な文字というカテゴリを IVS に入れてしまうことは，UCS 符号化文字を安定させつつ人文学研究上のニーズを満たすという点では意味があると考えられる．また，IVS とはいえ，標準の一環として定義され流通したなら，コードポイントは比較的安定したものとなるため，同系・非同系が曖昧である文字かどうかといったことをはじめとして様々な付帯情報をそれぞれの字形に対してアノテーションを付与して様々な解釈を記述していくことも，これまでの SAT のように独自の文字番号に基づいて行うことに比べたなら有用性は大きく高まるだろう．そして，ユーザレベルでの有用性についても，IVS 対応アプリケーションが今後順調に増えていったなら，独自の文字番号の利用とはまったく異なる次元のものとなるだろう．

2.6 SAT の外字に関する現在の取り組み

そのような見通しから，SAT では，UCS 符号化提案に向けて舵を切るに至った．情報企画調査会 SC2 委員会の多大なる協力を得て，現在は，漢字に関しては Extension F 提案の一部として IRG での議論が始まったところである．なお，SAT では，日本の代表とは別に，国際的な研究者グループとして IRG に参加し符号化提案を行っている．このように政府・地域代表以外のグループが IRG に直接提案をすることは初めてのことであり，今後はそのようにして研究者グループが直接符号化提案をするケースも出てくるであろうことから，以下，この符号化提案に際して活用したデジタル技術について報告することで，後に続く試みに少しでも役立てていただけたら幸いである．

3. 外字の符号化とデジタル技術

外字を符号化提案する際には，まず，その文字が本当に符号化されていないかどうかを確認することが重要であり，次に，すでに自分らのプロジェクトにおいて外字として登録されていないかを確認することである．

3.1 符号化文字と対応の確認の仕方

符号化文字として登録されているかどうかを確認するには，CHISE[l][12]と文字鏡検索が有益である．CHISE の場合には，IDS (Ideographic Description Sequence) と呼ばれる文字の部品文字列に対応する検索が可能となっており，探したい文字の一部の部品を入力して検索すればその部品が含まれる文字がリストされるようになっている．そこから，各文字の頁を表示させると，包摂となる文字についても一覧され，個々の文字間の関係も表示されるようになっている．また，唐代拓本文字データベース[m][13]に収録されている文字については，クリックすると唐代拓本文字の画像が列挙されるようになっているため，字形のバリエーションについての確認もしやすくなっている．一方，文字鏡検索の場合にも，部品による検索などが行えるようになっており，また，クライアントアプリケーションとしての便利さもある．なお，いずれも，すべての文字まで確認できるわけではなく，今回の場合には，特に Extension E の文字がまだ IRG 検討中の段階で対応が難しく，IRG が公開している文書などを追いかけて IDS や PDF ファイルを検索したりすることになった．

3.2 字形がはっきりしない場合の対応

大正蔵においては，活字であるにも関わらず，字形がはっきりと判別できないケースが散見される．また，初版と後の版で字形が変わっている場合がある．そのような場合には，まず，底本とされる高麗蔵が版本画像を Web で確認可能となっているため，これにアクセスして確認した．たとえば，Fig. 1 は，冠の上の点があるかないかの例であり，左が高麗蔵，右が大正蔵普及版となっている．そして，紙媒体上で確認した大正蔵の初版本では，高麗蔵と同様，上に点があるため，ここでは上に点をつけた文字とみなした．

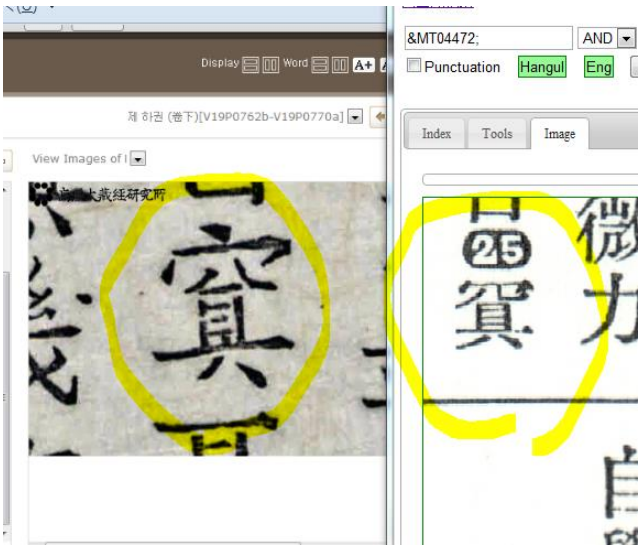

Fig. 1 高麗蔵（左）と大正蔵普及版（右）の例

あるいはまた，高麗蔵では十分な情報が得られない場合には，いったんデジタル媒体を離れ，前出の縮刷蔵を確認し，それでも判別できない場合には，宋・元・明代に発刊された木版大蔵経にて確認を行い，どのような文字として記述しようとしたのかを確認した．また，日本撰述部をはじめ，高麗蔵に収録されていない仏典，あるいは，古写経への参照等に関しては，可能な範囲で参照された一次資料にあたって確認を行った．そして，そのような文字の場合には，手書きで書かれた当時の字形となっていることが多く，それが現代の字形だとどのように記述し得るのかを確認する必要があるため，前出の唐代拓本文字データベース，及び HNG 漢字字体規範データベース[n][14][15]を参照し，確認

l) http://chise.zinbun.kyoto-u.ac.jp/
m) http://coe21.zinbun.kyoto-u.ac.jp/djvuchar
n) http://www.joao-roiz.jp/HNG/

を行うという形で再びデジタル媒体に戻ってきたのであった．いずれのデータベースも，それぞれの時代の字形の用例が簡単に検索・一覧できるようになっており，この種の用途には欠かせないものである．

3.3 外字情報の共有

外字を調査・確定していく作業においては，同じ文字の情報を登録してしまうことを避けるために，効率的に共有する仕組みが必要となる．望ましいのは，文字情報を登録する際に，すべての文字情報をチェックして同じ文字が登録されていないかどうかを確認でき，同じ文字があれば登録しないようする仕組みである．これが自動的にできるのであれば大変望ましいが，漢字の情報について既存かどうかを自動的に確認するのは容易ではなく，検索システム的なものを用意して人の目で確認するのが現実的である．特に，IDS をはじめとして，符号化されていない漢字を部品で検索する手法が近年は特に広まりつつあり，それなりに有効な手法となってきつつあるものの，文字情報登録の際に適切に部品を登録できるかどうか，登録された部品をうまく検索できるか，とったところにやや困難さが残っており，部品の階層性を適切に記述し検索できるようにすればいずれはかなりの精度で確認できるものと思われるが，現時点ではまだ最終的には人の目で確認する必要がある．そこで，検索できる仕組みを用意しておきつつ，いくつかの情報とともに文字情報を Web コラボレーションシステム上に登録するというのが現時点では現実的な解決策の一つだろう．SAT では，2005 年よりこのような Web 外字コラボレーションシステムを開発・運用してきており[16]，現在では 1 万文字を超える仏典に関する外字情報を収録している．

システムは，Linux, Apache, PostgreSQL, PHP を用いて開発されたものであり，SAT による他の Web コラボレーションシステムと同じサーバ上で運用されている．文字情報は，「合成表記」と呼んでいる独自の記述方法に基づいて文字を表現した上で，部首，総画数，内画数，音，類似漢字，その他情報からなっており，これに，文字情報記述者情報，文字画像作成者情報をそれぞれ記録し，文字情報記述は過去記録を保存することでトレーサビリティを確保し，クオリティを高めるために利用している．文字情報については，まとめて CSV 形式で一括ダウンロードできるようになっており，また，1 頁 200 文字分ずつで文字画像と文字番号のみを一覧できるようにもしている．文字画像に関しては，上述のように，GT 漢字のフォントを利用して作成した文字画像を主に用いている．また，さらに，今回の UCS 符号化提案にあわせて花園明朝フォントベースのフォントを GlyphWiki 上で作成中であり，このフォントから作成した文字画像へと徐々に入れ替えが行われているところである．

3.4 UCS 符号化提案文書作成にあたって

SAT では，いわゆるボーンデジタルな外字情報が集積されてきていたため，UCS 符号化提案にあたっては，技術的にはそれらの情報を規定のフォーマットにあわせて変換するだけでよく，そのためのプログラムの作成には多少の時間がかかったものの，資料作成そのものにはそれほどの労力をかける必要はなかった．提出を求められたのは，規定の書式に加えて，CSV，もしくはエクセルの符号化提案文字情報リスト，それから，それらの文字に関するエビデンスを収録した PDF ファイルであった．フォーマットも決まっており，IRG が定める”IRG Principles and Procedures”にその情報が掲載されている．文字情報リストに関してはさほど問題なく PostgreSQL に収録されたデータを変換するだけのことだったが，エビデンスの PDF ファイルに関しては，符号化提案文字の用例のコピーである必要があったので，それを新たに用意する必要があった．SAT では，すでに大正蔵テキスト部分 85 巻のすべての頁画像をスキャンして公開していたため，これを利用することができた．また，SAT のデジタルテキストには頁番号・段番号・行番号がすべて付与されているため，これを利用することで，頁画像上のどの位置に当該文字が登場するかということを，おおよそ把握することができた．すなわち，外字番号でテキスト検索をして登場箇所を特定し，その箇所の周辺の画像を頁画像から切り出すプログラムを作成することでエビデンスファイルに必要な画像を得ることができたのである．あとは，作業に関しては全般的に PHP を用いていたので，mPDF という PHP の PDF ライブラリを用いて自動的に PDF ファイルを作成した．当初の提案文字数は一回の提案文字数上限の制約等から 3514 文字となったので，3514 頁＋アルファの PDF ファイルを，用意されたデータから自動的に生成し，提出に至った．

4. 終わりに

一般的な Unicode のメリットと同様に，人文学研究にとっても，外字の UCS 符号化には利便性の向上や情報共有のしやすさ，国際化といった点で大きな意味があり，IVS の登場と普及の始まりは，そこにさらなる新たな可能性をもたらした．漢字文化圏，特に日本の人文学資料のデジタル化やデジタル技術を用いた応用的研究が欧米に比べて遅れをとっている大きな理由の一つは，人文学の学術的要請という観点からは文字が必ずしも適切に符号化できていないという点であった[o]．本来，コンピュータ上で文字はどう扱われるべきか，漢字はどう扱われるべきか，という問題は基礎的な問題として研究され議論され続ける必要があるが，一方で，今，そこにある資料をどのようにしてデジタル翻刻するかということはますます喫緊の問題として我々の前に立ちはだかりつつある．その問題を現実的に解決する一つの糸口となり得る道筋が用意されつつあることで，

o) 西欧語圏においても符号化されていない異体字の問題は存在するが，数はそれほど多くはなく，漢字文化圏とは状況が異なっている．[17]

日本の人文学研究も，ようやく本格的なデジタル翻刻へと進む見通しが立っていくかもしれない．そして，コンピュータ上での漢字の本来のあり方を問う議論を意識し，可能な限り反映させていくことができたなら，それは，さらに高い価値を持つものとなっていくことだろう．

大正蔵の UCS 符号化提案は情報規格調査会 SC2 委員会の多大な協力の下で実現し，今は IRG のメンバーのサポートも受けつつ進展している．大正蔵の文字情報の符号化は，単に産業振興のために工業標準を充実させるというだけでなく，文化的資産をデジタルの世界に適切に継承していくという意味合いを強く持っている．そして，その成果だけでなく，研究者グループが IRG に対して独自に符号化提案を行うというプロセス自体にも特徴がある．今回の大正蔵の UCS 符号化提案が様々な面でテストケースとなり，人文学研究におけるデジタル翻刻に新たな可能性を拓くことができれば幸いである．

なお，大正蔵には，UCS 符号化されていない文字群として梵字も多く含まれている．これに関しては，ちょうど時を同じくして，Anshuman Pandey 氏が WG2 に提案してきている[p]．ただし，この提案では，異体字に関しては，インド系文字の処理方法を適用しつつフォントを切り替えれば対応できるとされているが，専門家の協力を得て確認してみたところ，i や ii の母音の字形など，いくつかの字形に関して異体字を variant selector 等の形で登録する必要があると思われた．そのため，これも情報規格調査会 SC2 委員会と連携しつつ，異体字の扱いに関する追加提案文書を作成中である．具体的には，近々WG2 に文書が提出される予定であるのでそちらを参照されたい．

謝辞　本研究にあたっては，本文中に挙げた方々に加えて，多くの研究者の方々のお力添えをいただいたことを記して感謝したい．なお，本稿で報告されている成果には，JSPS 科研費 22242002 の助成を受けたもの，JSPS 科研費 60601680 の助成を受けたもの及び JSPS 科研費 30343429 の助成を受けたものが含まれている．

参考文献

1) Melissa Terras: Present, not voting: Digital Humanities in the Panopticon: closing plenary speech, Digital Humanities 2010, *Literary and Linguistic Computing*, Vol. 26, Issue 3, pp. 257-269.
2) 高楠順次郎編『大正新脩大藏經』大正新脩大藏經刊行会, 1924-1934.
3) 永崎研宣「大蔵経の歴史と現在」『新アジア仏教史 15　日本Ⅴ現代仏教の可能性』佼成出版社, 2011, pp. 15-53.
4) 『大日本校訂大蔵経』弘教書院, 1883-1885.
5) 李圭甲編『高麗大蔵経異体字典』, 高麗大蔵経研究所, 2000.
6) 山崎清華,「異字の撰擇に就いて」,『現代佛教』1928 年 11 月号, pp. 103-115.
7) 船山徹「漢語仏典―その初期の成立状況をめぐって」『漢籍はおもしろい』研文出版, 2008, pp. 71-118.
8) 三枝充悳『中論偈頌総覧』第三文明社, 1985.
9) 須永哲矢, 堤智昭, 高田智和「明治前期雑誌の異体漢字と文字コード」『人文科学とコンピュータシンポジウム論文集』Vol. 2011, No. 8, 2011, pp. 381-388.
10) 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』Vol. 2007, 2007, pp.237-244.
11) 小林龍生「区別することの必要姓を明確化することの必要姓ということについて」『人文科学とコンピュータシンポジウム論文集』Vol. 2011, No. 8, 2011, pp. 347-352.
12) 守岡知彦「文字オントロジーに基づく文字処理について」『情報処理学会研究報告』2006-CH-72, 2006, pp. 25-32.
13) 安岡孝一「拓本文字データベースの設計とその応用」石塚晴通編『漢字字體史研究』, 勉誠出版, 2012, pp. 116-128.
14) 石塚晴通，池田証寿，高田智和，岡墻裕剛，斎木正直「漢字字体規範データベース（HNG）の活用―漢字字体と文献の性格―」『人文科学とコンピュータシンポジウム論文集』Vol. 2011, No. 8, 2011, pp. 339-346.
15) 石塚晴通編『漢字字體史研究』, 勉誠出版, 2012.
16) 永崎研宣, 鈴木隆泰, 下田正弘,「大正新脩大藏經テキストデータベース構築のためのコラボレーションシステムの開発」『情報処理学会研究報告』CH-70(2006 年 5 月), pp. 33-40.
17) Medieval Unicode Font Initiative,　http://www.mufi.info/ (2012/12/21 閲覧)

p) http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4294.pdf