数理解析研究所講究録 1444 巻 2005 年 222-233

数学史文献資料の情報処理にまつわる諸課題について

東京学芸大学教育学部　渡辺　純成　(Junsei Watanabe)
Faculty of Education, Tokyo Gakugei University
junseiw@u-gakugei.ac.jp

0　はじめに

この研究集会の最終セッションでは，数学史文献資料の情報処理について，電子データの入出力と共有にまつわるさまざまな問題（具体的には，研究目的に応じた電子化のレベルの設定や，文字コードやデータフォーマットの選択，さらに研究成果の公表に際しての電子データの処理に伴う技術的・経済的な問題）に関する，情報交換が行なわれた．筆者が司会を勤めた．まず城地茂氏（高雄第一科技大）に，日中台の文字コードに関する現況と，中台で構築され利用されている，漢文文献の大規模な電子データベースに関する現況とを，報告していただいた(1)．つぎに，小川東氏（四日市大）から，和算文献の情報処理に関して，数学史研究者のあいだで共有可能な電子データを構築することへの必要性が提示され，さらに，その構築に際しての文字コードとデータフォーマットの選択に関する提言をいただいた(2)．ついで，自由討論のかたちで，技術的・経済的なさまざまな問題に関して，複数の方々から発言をいただいた(3)．

このセッションでは，現時点での諸問題を認識し，それらを巧拙はともかく何とか切り抜けるための初等的なノウハウを研究者の間で共有することに，力点があった．したがって，学問上または技術上，詳細な議事録を残すほどのものではなかったし，また，残してもいない．明記すべきことは，数学史に関する，数学史研究者のあいだで共有可能な電子データフォーマットを選択し，必要に応じてデータベースを構築し共有することの重要性と緊急性が，出席者の間で認識され，かつ，速やかに実現すべき課題として合意されたことである．この稿では，議事録に代えて，これら数学史文献資料の情報処理に関する基本的な問題点の極めて大まかな提示と情報源の紹介を行なう．情報源に焦点を当てた理由は２つある．１つには，多言語情報処理の世界では技術的細部があまりにも速く陳腐化するため，情報そのものよりも情報源を押さえるノウハウのほうが有用なためである．もう１つには，筆者は特にこれらの問題の専門家ではないので，読者が情報源に直接接触されるほうが適切であると考えたためである．

なお，上記の合意は，暗黙の前提として，数学史に関する文献資料の公開と共有が，オリジナリティをもつ研究よりも優先して，可能な限り速やかに行なわれるべきであるとの共通認識を含んでいることを，注意しておきたい．この認識は極めて正当なものである．論拠が，あらゆるひとに対して再検討可能な形態で公開されていなければ，学問ではない．いわゆる「前期旧石器」のスキャンダルも，その根源

は，一部の「研究者」が論拠を独占して公開しないままに，「成果」を積み上げたことにあった，ということができる．この立場からすれば，実は数学史文献資料について，校勘・注釈の作成よりもまず先に，良いテクストをそのまま影印して出版する，あるいは画像ファイルとして流通させることを通じて，研究者間で共有することを考えたほうが，はるかに適切である(4)．ただ，この場合には，法的(5)もしくは経済的(6)な問題は伴うものの，情報処理に関する技術的な問題はほとんどない．そこで，影印についての議論は省略する．

1 文字コードに関しては，日本のJIS，中国のGB，台湾のBig5と，そのさまざまなヴァージョンの収録字形数について．漢文文献の電子データベースに関しては，webサイトでは，台湾中央研究院の漢籍電子文献資料庫と故宮（寒泉）古典文献全文検索資料庫について．CD-ROMやDVD-ROMの形態で販売されているものでは，『文淵閣四庫全書』や『四部叢刊』とそれらの価格について．詳細については，この小文の最後の節で紹介する各種の情報源を参照されたい．

2 ユニコードを用いた単純なテクストファイルで保存し，割注などの形態情報についてはテクストに付けるタグのなかで記述するのが，最も汎用性が高い，と提言された．この提言に筆者も賛成する．また，和算文献の電子化に際して，算木記号の扱いやすい符号化も必要であることに，言及された．

3 和算文献の電子入力に際しての経験から，グリフ集合の規格やフォントへの実装に関する不満が述べられた．また，別の出席者から，近年，学術書の出版が極めて厳しい状況にあることにも不満が述べられた．いっぽう，何を議論しているのかよくわからない，という発言が聞かれたことも事実である．

4 清朝考証学の精度で校勘することには，多大の時間と労力と見識が必要とされる．したがって，文献資料に対してひとつひとつ校勘・注釈の作成を行なっていたならば，何時まで経っても資料は公開されないことになる．また，「へたな校勘よりもなまの資料を出してもらうほうが，学問上有益である」という辛辣な感想も，中国古典学の世界ではたびたび述べられてきた．

5 著作権が終了している古典籍であっても，影印に際しては所有者の許可が必要である．

6 紙媒体での出版では問題となる．画像ファイルの集積をCD-Rに焼いて流通させるだけならば，今日では，経済的にはさして問題ではない．

1 基本的な問題点

1.1. 文字とグリフとフォント

多言語情報処理に関して基本的な概念である，「文字」「グリフ」「フォント」をまず説明する．以下の説明は，

> ISO/IEC TR 15285:1998 "Information technology–An operational model for characters and glyphs"

で与えられた定義の，（三上，2002）(1)での翻訳と解説とに基づいている．

ISO/IEC TR 15285:1998では，「文字」「グリフ」「フォント」を以下のように定義する(2)．

> 文字：データを表現，構成または制御するために用いる要素の集合（文字集合）の構成単位，

グリフ：特定のデザインからは独立した，認識可能な抽象図形記号，
フォント：同じ基本デザインを持つグリフイメージの集合．

「フォント」ということばは，同じ基本デザインを持つグリフイメージの集合の特定の実装を指していうこともある．PostScript明朝体フォント，TrueType明朝体フォント，などというときの用法である．また，グリフイメージの集合の特定の実装形式を指していうこともある．PostScriptフォント，TrueTypeフォント，などというときの用法である．ISO/IEC TR 15285:1998での「フォント」は，むしろ「グリフ集合名＋書体」によって呼ぶほうが，適切かも知れない．

文字とグリフの関係は，アラビア文字で考えるのが最もわかりやすい．アラビア文字は，いわば続け書きの草書体しか存在しない文字体系であって，同じ文字であっても単語内での位置に応じてかたちが異なる．この異なるかたちが，グリフである．平仮名であっても，草書体で続け書きするときにはかたちが変化することを思い起こせば，文字とグリフがまったく異なる概念であることは理解できる．この場合，かたちだけで整理し分類したのでは，辞書式順序による並べ替えなどと整合しないため，抽象的な情報処理の単位としての「文字」を，具体的なかたちとしての「グリフ」とは別に，設定する必要があるわけである．漢字について，異体字を文字が等しいがグリフが違うものとして把握できるか否かには，問題が残る(3)．

フォントとは，各グリフの具体的な実現である２次元図形を，グリフ符号に対応させたものである(4)．テクストデータが画面に表示される際には，文字符号の有限列をグリフ符号の有限列に変換し，さらにグリフ符号にグリフ図形を対応させることを行なっている．最近の商業フォント（この「フォント」は，グリフイメージの集合の特定の実装を指している）は，フォント・ファイルの本体には，内部的な符号で順序付けられた図形を収めており，外部のグリフ符号を内部的な符号に対応させる部分は，単なる対応表で済ませている(5)．

あるコンピュータで出力可能な，具体的な図形としてのテクストは，その範囲が文字の符号化とグリフの符号化とフォントの実装のすべてに依存する．符号化された文字集合の大きさによって，コンピュータで処理が可能なテクストの範囲が，根本的に制限される．符号化されるグリフ集合の大きさによって，画面表示または印刷が可能な，具体的な図形としてのテクストの範囲が制限される．フォントの実装形態によって，画面表示または印刷の具体的な方式が決定される．問題点の所在は分離させる必要がある(6)．そして，文字の符号化とグリフの符号化の良否で決定される問題において，「フォント」ということばを用いることは，混乱を招く原因であるから避けたほうがよい．

中東系やインド系の諸文字の情報処理を適切に行なうためには，コンピュータのOSが文字とグリフを区別して扱うことができることが，必要である(7)．読者がOSを選択される際には，この点に留意されたい．

1　最後の節で紹介する文献については，以下，このように略記する．
2　機関によって定義が少しずつ異なる．注意されたい．
3　（情報処理学会文字コード委員会, 2002）と同委員会の議事録（ともに公開されダウンロー

ド可能である）で議論されている．

4 ここではグリフがフォントに先立つものとしてあるが，漢字の処理では，グリフを定めるには，やはりある標準的なフォントでの実現が必要なのではないか，という見解がある．漢字のある字形が規範性を獲得するのは，事実上，それが明朝体で表示されたときであることを思えば，この見解は穏当である．また，異体字に対しては，各フォントに共通するグリフ符号を与えることなく，各フォントごとに別個に処理することを許す，という案もあるが，これは最終的に標準的フォントの符号化が通行することになるであろうと思われる．これらの点について，（情報処理学会文字コード委員会, 2002）ではいろいろ議論されている．

5 グリフの符号化方式が，政府や企業の思惑でどうなろうとも，こうしておけば対応表を差し換えるだけで済む．

6 そして，責任の所在も．グリフ規格で規定されたグリフのすべてがフォントに実装されているとは限らないが，それは規格作成者の問題ではなく，フォント開発者の問題である．

7 もし区別して扱っていないならば，そのOSの開発者が如何に多言語処理を宣伝しようともそれは誇大宣伝に過ぎない，というのが筆者の意見である．このような開発者は，中東系やインド系の諸文字，そしてそれらによって記された膨大な文化と歴史の存在について，気付いていない．たとえば，蒙古文字・満洲文字で書かれたテクストの情報処理を差し置いて，トンパ文字の実装を誇るのは，開発者の見識の低さを表わしている．

1.2. 文字体系・グリフ体系の揺らぎに伴う問題

先に述べた「文字」「グリフ」「フォント」の定義は，情報処理の単位としての「文字」が，既に決定されていることを前提としている．しかしこの前提は，文字体系にあいまいな点がない，最近に作成されたテクストについてしか成り立っていない．どのような具体的文書にせよ，最初に与えられているのは，然るべき条件を満たす実２次元の有界図形たちの集合である．これらの図形を，平行移動やスケール変換などの無視を含むある同値関係で割り，その同値類の各々に対して代表元(1)を対応させるという操作で，グリフが定まる．さらに，それらグリフの集合に対して同値関係を入れて，その同値類の各々を文字とみなす(2)．ところでこれらの同値関係は，過去に作成された文書に対しては不明な場合がままある．また，同値関係が，時間に依存して変化する場合も多い．これらの不明さや不定さによって，具体的な図形としてのテクストをグリフ列や文字列に変換するしかたが，決定できなくなる．漢字圏では，具体的な図形としてのテクストを，抽象的な文字列として離散化しながら電子化するうえでの，原理的な，そして最大の困難は，同値関係のこの揺らぎである．

筆者の私見を述べる．これらの同値関係を考えるうえで重要なことは，グリフも文字も，体系をなしており，個々のグリフや文字は体系の中で始めて意味をもつ，ということである(3)．２つの図形は，それが１つの文字体系の中で意味の違いをもたらさなければ，同じ文字に対応する．異なる図形に見えたとしても，同時に出現し，かつ，異なる意味を担って使い分けられているのでなければ，文字体系の中の文字として区別する必要はない(4)．異なるグリフ体系に属する個々のグリフを直接比較することは無意味である．グリフ体系の組織的な対応関係を確立してから，体系の中での個々のグリフの機能を比較することにのみ，意味がある．したがって，

特定の時期と地域におけるグリフ体系と文字体系を，変種に到るまで明らかにし，さらにそれらの体系の時間的変化を明らかにしたうえで，必要な文字種をすべて含む最小の文字体系の中に埋め込むのが，原理的には最も適切なやりかたである(5)．グリフや文字の時間的変化は，体系間の関係として記述することにすれば(6)，グリフ集合や文字集合の無制限な膨張を防止することができる(7)．具体的な図形としてのテクストを文字列に変換するに際しては，まず，テクストが作成された時期と地域を特定し，そこで通用した文字体系とグリフ体系のすべてについて，変種レベルに到るまでの充分な知識を持ち，作成者がどの変種にしたがったのかを判断してから(8)，変換することになる(9)(10)．

漢籍の大規模な電子データベースとして成功を収めているものに，台湾中央研究院の漢籍電子文献資料庫があるが(11)，ここの『二十四史』データベースは定評のある北京中華書局の点校本を入力しており，グリフと文字の整理は入力前に印刷物の段階で終わっている．つまり，グリフ体系と文字体系について充分な理解がなされてから作成されている．

具体的な図形としてのテクストが採用するグリフ体系と文字体系について，充分な理解がないうちは，文字列への適切な変換は望めず，暫定的な変換に留まらざるを得ない(12)．具体的には，図形やグリフのあいだの同値関係を，使用目的に必要なところまで緩めたうえで(13)，与えられた図形としてのテクストを文字列に変換する．変換された文字列は，画像データと併用し，その索引として利用する，ということになる．与えられた図形としてのテクストがグリフ列に変換された場合には，文字を定めるグリフ集合上の同値関係も同時に与え，検索の際にはグリフではなくグリフの同値類を検索するように，電子データベース全体を設計するのが，適切である(14)．

1　前節の注１で触れたように，漢字のある字形が規範性を獲得するのは，それが明朝体で一意的に表示されたときである．草書体の漢字と簡体字の関係を思い起こされたい．

2　厳密にいえば，これは漢字の場合であって，中東系やインド系の諸文字の場合には，グリフ列の同値類を文字列とみなす，としかいえない．正確に述べるのはめんどうなので，このような言い方をするが，中東系やインド系の諸文字については適宜修正しつつ読まれたい．

3　言語の本来的にもつ体系性の反映である．

4　音声を例に取れば，現代日本語の発音体系では，ｒと１の違いは意味の違いを担わないため，音素として区別されず，表記もされない．それと同じことである．

5　通時的な資料群を情報処理するための文字符号やグリフ符号の設計にも，資料群の成立時期を含む時代と地域におけるグリフ体系と文字体系を詳細まで明らかにし，それらの時間変化を追跡することが必要である．幾つかの字書に含まれるすべての字形に文字符号を機械的に与えれば済む問題では，決してない．

6　巨大なグリフ集合の中の標準的グリフの変遷として，時間変化を記述することも，論理的には可能である．しかし，この場合には，グリフ集合の無制限な膨張を甘受することを迫られる．このやりかたは，具体的にいえば，略字や異体字の違いを，使用時期を無視してすべて保存して，電子化するということである．テクストが作成された時期と作成者を推定する場合に，略字や異体字の使用状況が手掛かりとなることは確かである．ただ，その場合には，画像そのものを保存してしまうほうが，筆跡や改変状況まで含み，はるかに情報量が多いわ

けで，手間をかけて文字列に還元する意義が乏しくなる．異体字の違いについては，同一テクストの中での使い分けがないならば，校記または解説での記述に留めるほうが，妥当であろう．

7 グリフ集合が膨張すると，１つのグリフに複数の符号が付されたり，具体的な図形を各グリフに振り分ける基準があいまいになったり，求めるグリフの所在が探し難くなったりする．大きいことは良いことである，とは言えない．

8 たとえば，あるテクストのある部分が略字や異体字を用いている場合には，作成者が自覚的にその略字や異体字を選んだ可能性もあり，正字を指示するつもりで略字や異体字を用いた可能性もある．後者の場合には，略字や異体字を保存することは作成者の意図に背くことであるから，略字や異体字を保存することがつねに正しいことであるとはいえない．したがって，テクストの性格に即しつつ，作成者が文字体系・グリフ体系のどの変種を採用するつもりであったのかを，判断しなくてはならない．

9 ここで述べた方針は，ことばづかいの新しさを除けば，漢人が経書の古文を隷書に書き換えたときの方針と，本質的には違いがない．なお，『論語』の価値は，そのテクストの字形に関わらず認められてきた．優れたテクストは，その書体や字形に無関係に貴ばれる．字形の選択においてしか独創性を主張できないような「文学作品」には，存在する意義がない．

10 ここでの議論は時間的変化についてのみ触れているけれども，実際には，空間的変化についても考える必要がある．異なる言語体系が，基本的に同一の文字体系・グリフ体系を採用しているけれども，これらの体系が完全には一致していない場合である．同一のグリフであっても文字としての意味が異なる場合には，翻訳処理などを考えて区別して扱うべきである（たとえば，「機」の簡体字と「机」の正字は区別するべきである）．しかし，かたちが極めて類似する２つのグリフが，それぞれ別の言語体系の表記において出現するが，同一の言語の表記において意味の違いを担って出現することがない場合には，これら２つのグリフを区別して符号化する必然性は，必ずしも存在しない．テクストが属する言語体系を指定すれば，これら２つのグリフの違いも指定できる．そして，スペルチェッカなどのテクスト処理を考えれば，テクストが属する言語体系を指定することはつねに有益かつ必要であるからでる．

11 URLは，大東文化大にあるミラーサイトのURLとともに最後の節で述べる．

12 したがって，基礎となったデータ自体が不変である場合にも，電子データベースを適切な状態に保つための保守作業が要求される．

13 字形自体の変異を研究する際には，既存のグリフ符号集合が小さすぎるという問題につねに遭遇する．この場合には，法的強制力を有する，またはもっとも広く実用に供されている外部のグリフ符号集合から出発し，それを拡張した内部的な符号で符号化し，外部のグリフ符号集合が拡張されしだい，内部符号を外部符号に順次置き換えていく，といった方針を取らざるを得ない．内部符号の設計については，符号の構造を１次元的にするか２次元的にするかなどで，いろいろな提案がなされている．（情報処理学会文字コード委員会, 2002）を参照されたい．また，内部符号の規格については，データの主たる利用者のコミュニティで合意されたものが望ましい．

14 さまざまなテクストデータに対してこの手続きが必要であるから，検索に際して個々のグリフではなくグリフの同値類を対象とするような手段は，OS側がもつことが望ましい．しかし，この手段の実装にはグリフ体系と文字体系について通時的な充分な理解が必要であるため，当分は実現不可能であろう．したがって，この手段は，現時点では個別のアプリケーションとして提供することを余儀なくされている．たとえばCD-ROM『文淵閣四庫全書』では，検索手段をデータベース付属のアプリケーションとして提供している．

1.3. 対応策

前小節で述べたように，現時点では，歴史的な文献について，異体字などの情報をすべて保存したまま電子化することは不可能である．グリフ体系と文字体系そのものを確定しなくてはならないのであるから，問題点が近い将来にすべて解決されることも期待できない．したがって，暫定的な対応に留まらざるを得ないが，この際に重要なことは，使用目的に応じて，データの電子化に必要な水準を冷静に検討してみることであろう．

使用目的としては，以下の場合が考えられる．典型的なものを挙げてみた．

(a) 電子的な索引として，特定の利用者のあいだで用いる．
(b) 電子的な索引として，不特定の利用者のあいだで用いる．
(c) コンピュータに処理・加工させて出力するための原材料として用いる．
(d) ひとがコンピュータの画面で読み取り，ときには加工するためのの原材料として用いる．
(e) 紙媒体に出力するための版下をコンピュータで出力する．あるいは，加工しない目的のPDFファイル，またはDVIファイルなどを出力する．

最後の(e)については，実は技術的な問題は，原理的には存在しない．版面を設計したうえで，不足する字形は画像として貼りつけるか，然るべき書体の外字として作成してデータに埋め込んでしまえばよいからである．法的[1]・経済的な問題が場合によっては生じるが，それらは技術的文脈で論じることではない．また，入力した電子データを紙媒体への出力にしか利用しないのは，労力の浪費であって，本来ならば(a)〜(d)としての利用がなされてしかるべきである[2]．したがって，(e)についてここでは議論を省略する．

(a)，(b)では，データフォーマットの可般性が高く，また，採用したグリフ・文字の符号化方式について，実際に広く利用されているものであることが要求される．特に(b)では，符号化方式の普遍性は必須である．利用者に別個にフォントなどをダウンロードさせ，インストールさせなければ使用できないデータベースは，まず使われないとみてよい．データフォーマットの形式としては，スタンドアローンならば単純なテクストファイルで充分である．その際には，付加情報は絞り込んだほうが使いやすい[3]．インターネットを通じて公開するならば，それなりのデータベース作成アプリケーションを用いて準備する必要がある．グリフ・文字の符号体系は，広く利用されている単独の体系を採用する．そこに収録されていないグリフ・文字は，規格書が紙媒体で広く行き渡っている体系[4]の符号を参照することで表現する[5]．利用者が特定の範囲に限られており，充分な知識を期待できる場合には，単に「〓」で表現しても支障は生じないものである[6]．なお，この場合には，異体字を横断して検索するしくみを提供することが必要である．

(c)では，データフォーマットに汎用性と可般性が要求される，処理する主体がコンピュータであるから，ひとの眼に関する可読性は必要とされない．したがって，タグ付きテクストを単純なテクストファイルとして保存したものが，最も適してい

る(7)．タグ付きテクストの形式については，情報技術一般に関する書籍や，この稿の最後の節で紹介する情報源を探されたい(8)．もっとも，既存の形式から外れていても，明確に定義され，機械的な検索と置換が行えるように設計されているならば，独自の形式を使っても構わないであろう．データを共有するために既存の形式に変換する必要が生じても，コンピュータによる機械的な作業で済むからである．グリフ・文字の符号体系は，広く利用されている単独の体系を採用し，そこに収録されていないグリフ・文字は，規格書が紙媒体で広く行き渡っている体系の符号を参照することで表現する．

(d)については，グリフ・文字の符号体系からはみ出るグリフ・文字の扱いに際して，厳格な符号化は必ずしも適切ではない．電子メールなどでこれらのはみ出る文字を扱う便法として，昔から，構成要素を列挙して「[ ]」などで括る，という方法がある．たとえば「草」は，「[艸/早]」もしくは「[艸+早]」などによって表現する．規格書を参照するよりも，速く理解できる．印刷出力するときには，フォントを指定できるアプリケーションに読み込ませたうえで，収録グリフ数が大きい文字体系のフォントのグリフに機械的に置き換えてから，出力する．

グリフ・文字の符号化方式について，「広く利用されている」という基準を述べた．このことについて，少し補足する．標準には，de jure 標準と de facto 標準の２種類がある．前者は，法的強制力を有する標準である．後者は，法的強制力を有しないものの，最も多くの利用者人口を有する標準である．両者が一致している場合には，それにしたがうのが最も問題が少なくて済む．一致しない場合，もしくは何れも存在しない場合には，見極めを要する．まず法的強制力が，de jure 標準の維持のための経済的支援を伴った，実効性をもつものであるか否かについて，注意すべきである．実効性をもたなければ，やがては de facto 標準が優越する．de facto 標準は，使用者の数が多いので，それが廃れ始めた場合でも新たな de facto 標準の規格にデータを変換する手段を，誰かが開発することが期待できる．この観点からすれば，ある規格は，その規格の使用者のなかに充分な能力と意思をもつグループが存在するならば，そのグループがデータ形式を変換する手段を開発するであろうから，選択肢として充分に有効である．最悪なのは，通達と勧告は出すものの，補助金と罰則は出さない（あるいは，出せない）機関が発表する規格である．この規格が主たる利用者から支持されていない場合には，なおさらである．

1 PDFファイルにフォントを埋め込む際には，フォントの使用許諾契約書の内容をまず確認してからに，されたい．この場合には，最終目的が，印刷用の版下作りなのか，あるいは電子的データの配付なのかで，許諾の可否が変わる場合がある．法的な面でも，使用目的を確定しておくことは必要である．

2 過去の文献資料の研究は，言語の分析から始めるべきである．言語の分析は，豊富な用例に立脚して進める必要がある．用例の検索は，ひとの眼で行なうよりも，コンピュータで行なうほうが，網羅的で効率的である．したがって，本文中の利用法 (a) は，過去の文献資料の研究において，その文献資料がある長さを超えるものならば，不可欠な段階である．ただし，文献資料の文字面や内容面での信頼性の評価は，少なくともいちどは資料全体を通読して，行なっておくべきである．

3 行番号は，原資料になくとも付けておくべきである．
4 unicode での符号，諸橋『大漢和辞典』の漢字番号，『今昔文字鏡』番号など．
5 公的機関の大規模な蔵書検索サービスの符号化方式については，利用した際の画面から想起されたい．
6 東洋文庫の漢籍蔵書検索では，JIS第２水準を超える漢字については「〓」で表現するが，混乱はない．
7 Tex，Adobe In Design，QuarkExpressなどのDTPソフトウェアでは，タグ付きテクストとして編集する，あるいは編集することができる．したがって，大量のテクストデータを出版する際にも，基本となるデータは，適切に設計されたタグ付きテクストとして保存しておくのが，賢明である．
8 研究機関が公開している電子データベース，具体的には古典籍の蔵書検索や，古文書の所在検索のサービスなどを提供しているものでは，作成方針を説明する部分で，タグ付きテクストの形式にまで言及している箇所が存在することも，ある．これらのサービスを利用される際には，このような情報にも注意されるとよい．実際に使用してみた経験を述べており，ときには，情報技術一般に関する書籍よりも有用である．また，情報の在り処がどの部局にあるかも推定できる．

2 文献とwebサイトの紹介

アジアの諸文字の符号化全般については，

三上喜連『文字符号の歴史－アジア編』，東京，共立出版，2002

が好著である．文字体系の分類（単子音文字，音節文字，結合音節文字，表語文字や非表音的記号など），文字の符号化の前段階としての各地での活字印刷の歴史から説き起こし，各文字体系の符号化の端緒と変遷について，図表を多用しながらていねいに解説している．それぞれのOSやアプリケーションの多言語処理の水準を測るための恰好の材料が，ここで提供されている．

東アジア諸言語の情報処理の実用的な面については，

Ken Lunde（著），小松章・逆井克己（訳）『日中韓越情報処理』，東京，オライリージャパン，2002

が有益である．日中韓越４カ国の言語の表記体系を始め，文字集合規格，符号化方式，入力方式，フォントフォーマット，出力環境，テクストの情報処理の具体的なテクニックなど，1000頁以上を費やして英語圏の技術者のために詳細に記述している．もっとも，原著の出版年が1998年であるため，最新の規格や最新のソフトウェア環境に関する記述は，当然存在しない．

JIS漢字規格の現況については，

芝野耕司（編）『増補改訂JIS漢字字典』，東京，日本規格協会，2002

が，JIS第４水準以下に収録された文字と記号について，典拠を示し，異体字への参照を載せて詳しい．コラムに示される編著者たちの見解について，筆者は必ずしも賛成するものではないけれども，JIS漢字規格への批判は，この字典の解説に示された文献学的作業と判断の水準を踏まえたうえで，しっかりした典拠と使用頻度の具

体的な数値を提示しながらなされるのが，建設的であると思われる．

日本の文字符号規格・グリフ符号規格のあるべき姿に関する議論は，webサイト

情報処理学会文字コード委員会
http://www.itscj.ipscj.or.jp/domestic/mojicode/index.html

にアクセスして，1999年８月と2002年３月の２回にわたって答申された，『文字コード標準体系検討委員会報告書』をダウンロードされたい．規格の作成・維持主体に関する議論から，日本の漢字の異体字の処理方式の具体的なアーキテクチャに関する複数の提案まで，多岐にわたる議論がなされている．このサイトにともに掲載されダウンロード可能になっている，『文字コード標準体系専門委員会議事録』（第１～第９回）とともに読めば，文字符号規格・グリフ符号規格とその実装に関するたいていの苦情は，本質的なかたちでは既に出尽くしているのがわかる．ここで論じられたことのその後の動きを観察すれば，速やかな解決を期待できることとできないことが，区別できる．

漢文文献の電子化については，2000年10月に創刊され，定期的に毎年１回刊行されている

漢字文献情報処理研究会『漢字文献情報処理研究』，東京，好文出版，

が，人文系の研究者にとって有益な情報を，多く含む．一般のソフトウェアや学術サイト・ソフトウェアに関するレビューが定期的に掲載され，また，電子化の具体的なノウハウや，電子化した先にあるべき学術研究としてのテクスト分析の具体的な事例が，掲載される．大陸中国の文字コードに関する最新情報は，印刷物の中では，この雑誌に掲載されるものが最も詳しいように，筆者には思われる．著作権に関する法的問題などが議論されることもあるので，漢文文献の電子化に関心をもつならば，この雑誌のバックナンバーには一通り眼を通してみるべきである．お手軽なハウツー本としては，この漢字文献情報処理研究会の関係者たちがやはり好文出版から，『電脳中国学』，『電脳国文学』などのシリーズを出している．ただしこれらの書籍は，その性格から言って賞味期限が相当に短いので，書誌情報をここに詳しく記す価値があるとは思われない．購入されたい方は，インターネットの書籍検索サービスで最も新しいヴァージョンを探すのがよい．漢字文献情報処理研究会のサイトは，

http://www.jaet.gr.jp/

にある．この会の中心的なメンバーのwebサイトも一見する価値がある．そのURLは，『漢字文献情報処理研究』の編集者・寄稿者の姓名で，インターネットを検索されたい．

電子データベース構築については，少々古いけれども，東京大学東洋文化研究所の紀要で「アジア研究とコンピュータ利用」特集と銘打った，

『東洋文化』No. 79, 1999年３月

に，現代中国書・和古書・インド学仏教学論文などの電子データベース作りで悪戦苦闘した当事者たちの経験が記されている．

中国学関連のwebサイトについては，改廃や内容の変更が激しいので(1)，ここで紹介に紙数を費やしても無駄になりかねない．読者の方々には，『漢字文献情報処理研究』のバックナンバーを検討されるか，あるいは

Yahoo! Chinese
http://chinese.yahoo.com/

で中国語圏のwebサイトを検索されることをお勧めする．ただし，Yahoo! Chineseは，簡体字中国語や繁体字中国語の入力と表示ができる環境であることが，要求される．日本のwebサイトについては，

Sinraptor Librarry of Sinology
http://www.ne.jp/asahi/sinology/lib/

のリンクが充実している．運営者は，中国文学の若い研究者である．

以下のwebサイトは，定番中の定番であるから，いちおう挙げておく．利用に際しては，繁体字中国語の入力と表示ができなければならない．

台湾中央研究院漢籍電子文献資料庫
http://www.sinica.edu.tw/~tdbproj/handy1/
大東文化大のミラーサイト
http://china.ic.daito.ac.jp/cgi-bin/handy/ftmsw3

『十三経』や『二十四史』などが，信頼のできる方法で入力されている．漢文をきちんと読む際には必ず典拠を明確にしなければならないが，その際に，この漢籍電子文献資料庫は，必ず当たっておかなくてはならない電子データベースである．OSとブラウザのヴァージョンによっては，大東文化大のミラーサイトのほうが問題が少ない．何れも特殊な外字を使用しているので，場合によっては字をハードコピーで確認することが必要になる．

故宮（寒泉）古典文献全文検索資料庫
http://210.69.170.100/s25/

も挙げておく．ここでは『全唐詩』『朱子語類』『四庫提要』などが利用できる．『二十四史』は削除された．

大陸中国にも有用なwebサイトの電子データベースがあるが，要求されるソフトウェア環境や使用料金の支払い方法など，いろいろな問題があるので，それらについての解説とともに『漢字文献情報処理研究』で調べていただくことにして，ここでは省略する．

CD-ROM版『文淵閣四庫全書』など，CD-ROMやDVD-ROMの形態をとって大陸中国の企業から販売されている電子データベースのリストと価格などの詳細については，国内各地の中国書籍取り扱い店に問い合わせられたい．

1　中国と台湾がWTOに加盟した際に，それまで電子データベースを公開していた複数のweb
サイトが，消滅するか，あるいは内容を削減した．著作権の強化に絡んだ動きと思われ
る．．