chapter.11

電子テキストの有効利用に関する雑感

――文献資料のモデル構築の可能性

宮崎 泉

1. はじめに

仏教学の分野では PC の利用が比較的早くから進んだ。文献を主たる資料とする研究分野では、多数の電子テキストの横断検索がとりわけ有用であったのが主な理由と思われる。以来、電子テキストの有無自体が利便性を極端に左右するため、とにかく無数の電子テキストが作成された。しかし実際には電子テキストの作成方法について何らかの基準や合意があったわけではない。それぞれが検索しやすい形でそれぞれに入力してきたというのが実情である。そこに入力ミスが残ることがあるのはやむを得ないとしても、逆に、入力者の「善意」でテキストが暗黙のうちに訂正（変更）されたりしたようなものも多いと思われる。その際、検索の利便性が最大の目的になってしまい、自身の入力しているテキストが一体何なのか、また、この分野の研究対象とどのような関係にあるのか、といったことはほとんど意識されてこなかったのではなかろうか。

近年これまで蓄積されてきた電子テキストをより有効活用する可能性について議論されるようになってきた。そのため、この時点で一度研究対象と電子テキストとの関係について検討しておくことは必要なことと思われる。もしその作業を通して今後何らかの合意が得られるようなことがあれば、将来的にも有用であろう。筆者にはこれまでにも何度かそのようなことを考える機会があったが、今回自身の考えを公にできる場を得たため、これまで考えていたことを再検討しまとめておきたいと考えた次第である。もっとも筆者に現在確たる結論があるわけではない。しかしながら、「テキスト」とは何かについて、実際

の研究対象にかかわる諸問題を取り上げつつ、電子テキストを扱うときに何が問題になるのかを考える材料を提供できればというのが本章の目的である。

2. 研究対象としてのテキスト

筆者はこれまで主にインド仏教の文献学的研究に取り組んできた。この分野では、サンスクリットなどのインド系諸語で書かれた原典が残っていれば、それが最も重要な資料になる。そのほかに、漢訳やチベット語訳が現存することもある。漢訳やチベット語訳は、ほとんどインド諸語から直接翻訳されたものであり、まとまった形で伝わっているため、インド仏教研究者にとって貴重な資料である。しかも、サンスクリット原典が失われていれば、漢訳やチベット語訳（一方しか現存しない場合も多い）だけでインド仏教を研究することになり、その意味でも不可欠の資料と言える。

さて、ここでこの分野の研究を具体的に紹介するために、サンスクリット原典、漢訳、チベット語訳が残っている場合を考えてみよう。このような研究分野に詳しくない人は、サンスクリットで書かれたオリジナルと、その漢訳者の理解と、チベット語訳者の理解の三つを知ることができると考えるかもしれない。しかしことはそう単純ではない。というのは、どれももとの形のまま現在まで伝わっているわけではないからである。サンスクリットは写本の形で伝わるが、写本が一つあるいは複数伝わっていたとしても、そのいずれもオリジナルのままであることはない。写本が伝わる過程でさまざまな要因によってテキストが変容するからである。複数の写本が伝わる場合、それらはいくつかの系統に分類できることがある。系統の分岐や系統の中の分岐も想定できれば、系統樹に表すことができる。しかし、写本すべてが同程度の価値を持つわけではない。例えば、ある写本が別の写本のコピーであることが明らかであれば、その写本の持つ価値がいちじるしく減少するのは明らかだろう。このような状況の中で、各写本の重要性も加味しながら比較検討を行い、なるべくオリジナルに近づこうと努力するのがこの分野の研究の出発点である。

ただし、その際、現存資料の最も古い分岐より前の情報は失われているため、オリジナルには決して到達できないことには注意しなければならない。新たに

より重要な写本が発見されれば、現在想定されている批判校訂テキストも大きく書き換わる。そのため、このオリジナルに近づこうとする文献学的研究に終わりはない、とも言えよう。サンスクリット写本以外にもオリジナルに近づくためのさまざまな手段を駆使しながら真実に接近し続ける営みが絶えることなく続くのである。ある文句が別のテキストにほぼ同じ形で現れる平行句が見つかる場合もあるし、対論者が批判のために引用することもある。翻訳もそのような情報の一つに数えることができる。思想や歴史研究の基盤として、このような永遠にオリジナルに近づき続けようとする文献研究がある。

では、漢訳やチベット語訳はそのようなサンスクリット原典とどのような関係にあるだろうか。上記のようなサンスクリット写本の伝承を考えれば、漢訳やチベット語訳がサンスクリットのオリジナルの翻訳ではあり得ないのは明らかだろう。漢訳やチベット語訳はそれぞれ、サンスクリットのある時点の一つないし複数の伝承を伝える。漢訳とチベット語訳の両方が残っていても、それは同じサンスクリットからの翻訳ではない。漢訳やチベット語訳の訳者にとって、目の前にあったはずのサンスクリット原典の状況はわれわれとまったく同じである。つまり、その時点で伝承されていたサンスクリット写本が一つないし複数あり、それをもとに翻訳が作成されたということである。もちろんその翻訳は、サンスクリット写本そのままが翻訳されたわけではない。サンスクリット写本の質が悪ければ、多くが訳者の知識によって補われ、訂正された後に翻訳されたであろう。また、関連文献の翻訳がすでにあれば、それも参照されたかもしれない。翻訳は、翻訳者の理解がサンスクリット原典の理解に資する場合も多いが、実際には翻訳の成立過程は複雑で、それに注意を払いつつ慎重に扱う必要がある[*1]。

しかも、翻訳も伝承される間に変容しながら現在に伝わっているのである[*2]。現在伝わる翻訳が、翻訳された当時のままであることはあり得ない。文献研究者は、このような状況にある資料からオリジナルのサンスクリットを再構成することを目指し批判的校訂を行う。しかし前述の通り、それでもオリジナルそのものには到達し得ないことには注意が必要である。写本に限らず、ある変容をこうむって以降の資料しか伝わっていなければ、結局確実な証拠をもってそれ以前にさかのぼることができない。以上を言い換えれば、失われたオリジナ

ルを基点に、それぞれ異なった性格を持つ多種多様のテキストが存在しているといえる。これが、われわれが一つの「テキスト」と呼んでいるものの内実である。

さて、ここまでは、たとえそこに到達できないとしても、オリジナルが一つだけ存在する前提の話である。しかし、実際にはオリジナルが変化することもある。例えば、ある経典に改変や増広が繰り返され、変容していくような場合である。このように変容していく全体を一つの「テキスト」と見なすとすると、オリジナルに一定の幅を許容することになるだろう。その場合テキストは文字そのものを離れ、より抽象的なテキストが想定されていることになる。さらに、研究をしていく上では、このようなテキストと別のテキストを一つの文献群として扱いたい場合もあるだろう。そのときには、文献群をさらにより抽象的なテキストとして想定していると考えることもできる。つまり、どの範囲を一つのテキストとして扱うかが必要に応じて変わるのではないかということである。

実例としては、般若経が最適である。般若経とは大乗経典の初期から現れる経典である。中でも『般若心経』はそのサイズの手頃さもあって、日本で最もよく知られた経典と言えよう。般若経は単一の経典ではなく、経典群であり、いくつかの系統が想定できる。最古の系統は「小品系」と呼ばれ、『八千頌般若経』の系統である。それに遅れるもう一つの重要な系統が「大品系」と呼ばれる『二万五千頌般若経』の系統で、八千頌より長いものである。そのほかに、それらよりはるかに小さな『般若心経』『金剛般若経』、逆に最も大きな『十万頌般若経』なども伝わる。それぞれの発展段階は、漢訳によく保存されている。例えば、「小品系」の最古のものは漢訳に『道行般若経』として伝わるほか、複数の漢訳が伝わり、それらは現存のサンスクリットよりも古い段階の小品系般若経の姿を伝えている。ここで一つの系統とみなしている小品系般若経は、オリジナルが変容しながら伝わる、広い意味でのテキストと言える。そしてまた、この般若経典群全体を一体として扱うなら、それはさらに広い意味でのテキストということになるだろう。

このように考えると、実際の研究の場では、テキストとは何かより、何を一つのテキストとして扱うかが重要であることに気づく。その場合、いずれにせよ、テキストはある程度抽象的なもので、ある観点から一つのテキストとして

想定されたものであると言える。逆に言えば、一つのオリジナルを想定しているような場合でも、その観点から複数の写本の背後に一つのテキストを想定しているのであって、そのこと自体は必ずしも自明ではない。研究の中では、それ自体が問われることもあり得る。写本の断片を扱っている場合などがわかりやすい例になるだろう。ともあれ、一つのテキストを区別する観点は無数に想定できるが、実際の研究の場でもそれを場合に応じて使い分けていることに自覚的になり、電子テキスト上でその観点を自在に切り替えられるようになれば、新たな地平が開けるのではなかろうか。次に、この複数の観点をもう少し噛み砕いて説明するために、いくつか別の話を紹介しよう。

3. テキストと漢字

では、そのような「テキスト」を実際に電子化するときにはどのように扱えばよいだろうか。最も単純なところから考えれば、テキストとは文字列の集合である。しかしそれは単なる文字の羅列ではなく、全体として意味を有する。むしろ、実際には人はその意味を扱いたいにもかかわらず、意味そのものを扱うことができないため、意味を文字列の集合に託しているにすぎない。コンピューター上でテキストを扱う場合も同じことである。さて、テキストが文字列の集合という形を取り、意味という内実を持つと考えるとき、それが「漢字」に非常に似ていることに気づく。そのため、より複雑なテキストについて考える前に、漢字がコンピューター上でどのように扱われているのかを見ておくことは有用であるように思われる。そこでまず「漢字」について考えるところからはじめてみよう。

現在のコンピューターの内部では、文字であれ、画像であれ、映像であれ、すべてが数字によって表されている。このことにいまさら説明の必要はないだろう。例えば、アルファベットであれば、アルファベット一つ一つに順に数字が割り振られている。すると、その数字の順に並べるだけで、アルファベット順に並べ替えることができるようにもなる。この数字は文字コードと呼ばれる。ラテン文字も文字の体系によってさまざまなので単純ではないが、テキストの問題について考える上では意味を有する漢字の方が参考になる点が多く、われ

われにとってより身近でもある。そこで、ここではラテン文字より複雑な漢字について取り上げよう。

漢字に文字コードを割り当てる場合にまず問題になるのは、その文字コードを割り当てる「ひとつの漢字」をどう考えるのか、ということである。文字は書かれたり、印刷されたりするもので、形を持つ。もちろん同じ形を持っている漢字が一つの漢字である。しかし、手書きされる場合には、多少はねやはらいが違っても同じ文字として通用するだろう。さらに、本来ないはずの点が打たれているようなこともよくある。そのような違いによってほかの文字と混同されない限り、同じ「ひとつの文字」として使用されているのが実際である。もう一方には、小学校ではねやはらいを厳しく注意されたように、規範的な文字の存在を想定することもできるが、漢字の現実はより緩やかで柔軟なものである。また逆に、ほぼ同じ形を持っているにもかかわらず、意味が異なる漢字もある。例えば、「干」や「于」は本書では「千」とは形が異なり、逆に書けばもちろん誤りと見なされるが、手書きであればこれらが区別のつかない形で書かれることもあるだろう。しかし、この場合は形がどれほど似ていたとしても「ひとつの漢字」と見なしてはならず、別の文字コードを割り当てる必要がある。意味が異なるからである。

文化庁の Web サイトに、平成 28 年に文化審議会国語分科会から出された『常用漢字表の字体・字形に関する指針』(以下、『指針』) と題する報告がある [*3]。『指針』は、コンピューター上での漢字の扱いに関する問題も踏まえて、手書きの字形やフォントのデザインについても詳細に解説しながら、常用漢字について説明しているが、その際漢字は「字種」「字体」「字形」のレベルに区別される。それぞれの用語の意味は「第 3 章　字体・字形についての Q&A」の中に簡単にまとめられているので、それを引用しながら紹介しよう。まず「字体」とは「文字を文字として成り立たせている骨組み」であり、「字形」は「それを実際に文字として記した時の形状」と言う（p.69, Q5）。字形は実際に目に見える多様な形なので想像しやすい。一方、字体はある文字を別な文字と区別するもので、より抽象的なものと言える。一般に字体という語は明朝体やゴシック体などの意味で使われることもあるが、『指針』ではそれは「書体」と呼ばれ区別されている。「字種」は「同じ読み方、同じ意味で使われる漢字の集まり（グルー

プ）を指す常用漢字表の用語」である（p.69, Q6）。例として「桜」と「櫻」が挙がる。いわゆる「異体字」を包摂して一つの文字と見なすレベルと考えることができるだろう。つまり、いくつかの字体の集合を別の字体の集合から区別するものであり、字体よりさらに抽象度が高いものだと言える。

文字コードの問題に戻れば、実際問題としては字体レベルで文字コードを割り当てることが多いと思われる。字形は無数にあって、その一つ一つに文字コードを割り当てることは不可能なので、字形レベルの多様性は包摂せざるを得ない。しかし、「桜」と「櫻」を同じ文字として扱うよりも、字体レベルの差異は区別したい場面が多いのだろう。別の文字コードが割り当てられた文字を同一視するのはたやすいが、同じ文字コードを割り当ててそれを区別するのは難しい。JIS 規格で定められた文字コード（例えば、JIS X0208 など）でも、「桜」と「櫻」は別の文字コードが割り当てられている。これは「字体」レベルで一つの文字と考えているということになる。ただし、テキストの検討の中にこのような考え方を導入するときは、一つのテキストを何か一つのコードに割り当てるなどということを考える必要がないため、ここで何を一つの文字と考えるべきかを論じる必要はない。むしろ、レベルが変わると、ある二つの文字が異なる文字と見なされたり、同じ文字と見なされたりする点が重要であり、それこそがここで注目すべき点である。

テキストに話を戻そう。いまの漢字の例をどのようにテキストに対応させるかにはもちろんいくつか可能性があるが、何よりはっきりしているのは、テキストが実際に書かれて伝わった写本が「字形」に対応するということである。文字の場合、目の前にある具象として存在するものは「字形」である。テキストの場合に、具象として存在するのは写本である。その二つを対応させることに疑問の余地はないだろう。では、そのような写本を一本ないし複数利用して批判校訂した校訂テキストは何にあたるだろうか。これは誤解されることも多いのではないかと思われるが、実際目の前にある具象という点で写本と同じであり、「字形」が対応する。現存するいくつかの「字形」を利用し、現代の学者が客観的によりよい「字形」を世に提示したものが校訂テキストであると言えばわかりやすいだろうか。そしてこのときには、暗黙のうちに複数の写本が一つのテキストであることが想定されている。そこで想定されている抽象的な

テキストが「字体」に対応すると言えよう。もっともすでに述べた通り、一つのテキストの内容にある程度幅があるときもあるが、それも大きな問題ではない。それは複数の写本を一つに考えるときにどのような観点を想定するかによる。すると、それより抽象度の高い字種に対応するのは、字体に対応させたものより抽象度の高いレベルのものとしか言えないことになるだろう。例えば、同じ写本の複数の系統を束ねるものとすることもできるだろうし、また、異本や翻訳も含めて一つにまとめたものを想定することもできる。さらに、文献群を一つの「テキスト」と考えるようなものをより上位に考えることもできるだろう。

このように考えると、テキストの場合『指針』にあげられているような三つの段階に単純に対応させることはできないことになる。目の前にあるものが字体に対応し、それが出発点であるのは明らかだが、それより上にはさまざまなレベルを想定することができるからである。もっとも文字の場合もさらに多様なレベルを想定することができるが、『指針』は常用漢字の考え方を説明するために、比較的わかりやすいモデルに落とし込み、うまく字形や字体を説明しているとも言えるだろう。

文字に対して多様なレベルを許容しそれを自由に扱うことを目指したシステムに、守岡知彦（もりおかともひこ）氏（京都大学・人文科学研究所）が主導してきたCHISE[*4]がある。CHISEでは、文字をオブジェクトとして扱い、文字コードとして符号化することのない文字処理の実現を目指した。そのモデルはChaonモデル[*5]と命名され、要するに、文字を文字コードではなく、「文字素性」（character feature）の集合として扱うものである。文字素性には、部首、画数、発音、意味のほか、ある文字集合の中のコードポイントなども含み、既存の文字コードを使用したシステムとの情報交換も可能にしている。さらに、その文字オブジェクトが継承関係を持つことも考えられており、それによって、あるレベルでは同じ文字だが、レベルが変われば別な文字になるようなものも扱えるようになっている。CHISEでは理論的には無数の階層を持つことができるはずであるが、記述が複雑になりすぎたり、利用者の直観に合わないケースが多発しかねないことを怖れ、ガイドラインを設けてある程度整理している。それが「CHISE文字オントロジーのための漢字字体・字形粒度の情報記述に関するガイドライン」[*6]で

ある。そこでは包摂粒度の観点から「抽象文字粒度」「統合文字粒度」「抽象字体粒度」「詳細字体粒度」「抽象字形粒度」「例示字形粒度」を設定し、常用漢字表の『指針』と比較するとはるかに詳細になっている。その際、守岡氏は前掲常用漢字表で用いられる「字種」—「字体」—「字形」という概念と、UCSなどで用いられている『抽象文字』—『グリフ』—『グリフイメージ』という概念が、おおまかに「字種」—「抽象文字」—「字体」≒「グリフ」—「字形」≒「グリフイメージ」という四階層の包摂粒度に整理することができると述べた上で、実際にはその四階層に整理できない場合が出ることを踏まえ、上記の階層を提案している[*7]。

ここではその詳細は省略するが、筆者が複数の階層を持つ抽象的なテキストのモデルを考えるとき、実はこの CHISE のモデルから多大な影響を受け、それが出発点になっている。つまり漢字の場合と同じように、テキストは文字からなるが、文字の集合そのものがテキストではなく、抽象的なテキストを想定し、文字の集合をその属性と考えることができないかということである。もちろん抽象的なテキストは、複数の文字の集合（文字としてのテキスト）を属性として持つこともできるだろう。また、抽象度を変えれば、いわゆる文献群も一つのテキストとして扱えるようにできるのではないだろうか。

4. テキストと書誌情報

ところで、図書館では具象としてのテキストを扱うが、書誌情報を考える場合には抽象的なレベルを想定することも必要になるようである。筆者はまったくの門外漢であるが、ここでの議論に参考になるところもあるため、2017年に改訂された IFLA LRM[*8] を取り上げ、われわれの分野における問題点を考えてみよう。IFLA LRM とは、国際図書館連盟（IFLA）[*9] が発表した書誌情報扱うための概念モデル（Library Reference Model）である。そこではテキストにいくつかのレベルが想定され、Work, Expression, Manifestation, Item の四つが定義される。Work が最上位概念であり、Item が一つ一つの書籍にあたる。Expression は特定の版や翻訳など表現によって区別されるものであり、Manifestation はその Expression がある出版社から刊行されたりしたもので、幾

分具体的なものである。Manifestation の例に ISBN が付いたものが挙がっていると言えば、わかりやすいだろうか。

このモデルに沿ってわれわれの分野を考えてみると、一つの経典あるいは論書が Work である。サンスクリット原典、漢訳、チベット語訳などが Expression にあたる。Expression はそれぞれのテキストにあたるが、一つに固定したものではない。それが写本や版本になったときに、Manifestation になる。版本の版の違いや写本の一つ一つはそれぞれ Manifestation である。写本の場合は Item は Manifestation そのものであるが、版本の場合は、一つの Manifestation の Item が複数の場所に存在することもあるだろう。また、それらの資料を利用して作成された批判校訂本の一つ一つは Manifestation である。それが改訂され版が異なれば、また別の Manifestation になる。

この中では Work が特に興味深い。Work は純粋に抽象的な概念であり、複数の Expression を束ねるものである。そして、それ自体いかなる具体的な存在とも同一視することが許されない。この考え方は、われわれが上に考えてきた中でも最も抽象的なテキストということになり、抽象的なテキストをどう考えるべきか、示唆を与えてくれる。もう少し要点を示せば[*10]、Work は、最初の Expression が完成したときに同時に生じる。一方で、すでに存在しているテキストから Work が考えられることももちろんある。その際、何が一つの Work かの判断は利用者のニーズが基盤となり、多くの利用者が一般的に一つの Work と考えるものが一つの Work である、と言う。多様な文化を背景にさまざまな形で成立した Work を厳密に定義することは不可能なので、運用で工夫する余地を残しているものと考えられる。この点もわれわれにとって非常に示唆的であるが、同時に、利用者に考慮の余地を残すと多様性を許容することになり、書誌情報を扱う場合には頭痛の種になるかもしれない。

さて、われわれの研究分野でも Work を狭い範囲で考えることができる場合は、LRM のモデルを適用することができそうである。具体的には、ある論師が著した論書を扱う場合などである。一方で、経典が発展しつつ伝承されるような場合は、その経典群は一つの Work と考えられるだろうか。また、般若経典などのように、内容的には密接に関係しつつ長さが異なる経典群を一括して考えたいような場合はどうすればよいだろうか。IFLA LRM では、Expression は、

別の Expression の aggregate とすることもできるため、それを利用することができるかもしれない。一方で、そのほかにも多くの relationship が定義されており、例えば、Work と Expression の関係は、is realized through によって表現される。また、Work 同士の関係は、has part, precedes などのほか、is inspired by, is a transformation of のような関係も想定されており、多様な記述が可能のようである。しかし、Expression 同士の関係を記述することはあまり考えられていないように見える。書誌情報を扱うという点からは当たり前のことだが、Work 中心のモデルが考えられている。けれども、われわれの研究分野のように、一つの Work の中の文献の状況が問題になる場合は、むしろ前述の Expression 同士の関係が問題になる。IFLA LRM でも Work を介して Expression 同士の関係を示すことは不可能ではないが、われわれの用途に利用するには少々改変する必要がありそうである。さらにその際、文献群のような、広い意味での Work を扱う仕組みも考えておいた方がよいように思われる。

われわれの分野と同じように、漢籍でもさまざまな問題が起こるらしい。図書館情報学を専門とする木村麻衣子氏（慶應義塾大）が「東洋学へのコンピュータ利用第 30 回研究セミナー」（2019 年 3 月 8 日、京都大学人文科学研究所）の中で、特に Work と Aggregate を取り上げ、漢籍の中で利用者がおおむね合意できる「著作」（Work）を決めることの難しさを指摘したことがある [*11]。ここでは詳細を紹介することはできないが、一つには曖昧な「著作」を想定せざるを得ない場合があるからだという。これは、IFLA LRM が Work に検討の余地を残しているからこそ生じる問題ではなかろうか。実際には利用者の中でも何が Work か一致しない場合もあるだろうし、研究者の場合はなおさらである。とすれば、われわれの分野でも IFLA LRM のモデルをそのまま用いるのは難しいように思われる。

このように、書誌情報を扱う IFLA LRM でも、抽象的な上位概念を想定した上で、一つの作品をいくつかの階層に分解してモデル化しており、われわれが文献研究に用いるモデルを考える上でも参考になることが多い。一方で、Work に曖昧さを認めながら、使用するときにはやはりそれを決定しておく必要があることなど、それ自体を問題としたい研究者にとっては、そのまま流用することのできないこともある。そのような問題点を精査した上で、われわれ

の実際の研究状況に合わせて、柔軟に抽象度を変更して扱えるようなモデルを考える必要があるだろう。

5. おわりに

このように、われわれこの分野の研究者は手に取ったり目にしたりできるテキストを扱いながら、実際には意識的にであれ無意識にであれ、より抽象的なテキストを背後に想定しつつ研究している。校訂テキストは、一度抽象的なテキストを経由し、それがより合理的な具象として改めて現れ出た具体的な成果と言える。だとすれば、具体的なテキストといくつかの階層を持った抽象的なテキストを考えて、それをモデル化すればよいように思われるかもしれない。しかし、研究者にとっては、何を一つのテキストと考えるか自体が研究対象であり、議論の対象である。また、研究の過程では、さまざまな形で想定した階層を行き来しながら研究を進めることもある。そのため、固定的なモデルを作ってしまうと、結局研究者は誰も使えないものになりかねない。とすると、どのようなモデルを考えるとしても具体的な一つ一つのテキストが出発点となるのは間違いないが、抽象的な階層は柔軟に変更できるようにしておき、それらの間の関係を記述できるようなことを考えるのが最も効率的なのではないだろうか。研究に使うためのテキストのモデルは完全なものである必要はなく、作業用の仮説的なモデルで十分のはずである。

最後に、冒頭にあげた電子テキストの問題にもう一度触れておこう。もしある電子テキストが既存の校訂や写本をそのまま入力したものであれば、それは既存の校訂や写本と同じ具体的なテキストである。しかし、誰かが誤りも含むテキストを作成したらどうだろう。その場合はもちろんその人が新たに別な具体的なテキストを作り出したことになってしまう。あるいは、誰かが何か校訂や写本の誤りを見つけたと考えて入力時に訂正したらどうだろう。訂正しておいた方が検索のためを考えても好ましく、多くの人がそうするかもしれない。しかし、その場合もやはりその人は新たに別な具体的なテキストを作り出したことになるのである。そのことを意識して行うのならまだよいが、実際には無自覚に行われることが多いのではなかろうか。

もっともこれはまさに写本の伝承の中で常に起こってきたことである。そう考えると、人間の営みとして自然で避けがたいことなのかもしれない。あるいは、校訂テキストでも新たな具体的テキストを作り出しているのではないかと言うならば、その通りである。しかし校訂テキストの場合は、校訂の方針を明示した上で使用した写本の異読も示し自覚的に新たな具体的テキストを作成しているため、現代の研究者にとっては問題ないのである。続く研究者がそのようなものとして意識的に扱うことができるからである。電子テキストの場合も、XML などを使えば、複数の写本の情報や入力者の訂正情報を示すことができるようにすでになっている。この時代に単純なテキストファイルを使って信頼度不明の具体的テキストをさらに作り出す必要はまったくない。テキストファイル形式で情報が欲しければ、そのような形の XML から切り出すこともできる。これからは、これまで蓄積されてきた大量の電子テキストをより信頼できる形に変えていくことが、今後電子テキストをより有効利用するための出発点となるだろう。上のようなモデルを意識しつつ、写本や版本の情報を正確に再現し、校訂や訂正はそれとは別に取り出せるような形の電子テキストを実現できれば、よりよいテキストに近づいていくための基盤となるはずである。

注

1　例えば、漢訳の成立と関連する諸問題については、船山徹『仏典はどう漢訳されたのか――スートラが経典になるとき――』（岩波書店 , 2013）を参照されたい。

2　チベット語訳では、伝承の過程で校正の際に改めてサンスクリット写本が使用されたことが奥付（コロフォン）に明示されていることがよくある。次にあげるのは極端な例で、サンスクリット写本三本と校合したというものである。これは著名なナーガールジュナ（龍樹）の著作『宝行王正論』（Ratnāvalī, D.4158, P.5658）チベット語訳の奥付であり、ここで引用したのはデルゲ版チベット大蔵経による。「……再びインドの学者カナカヴァルマンとチベットの翻訳官パツァプ・ニマタクがインドの写本三本と校合して修正した。」（rgya gar gyi mkhan po dzñā na garbha daṅ / bod kyi lo tsā ba dge sloṅ klu’i rgyal mtshan gyis bsgyur ciṅ źus te gtan la phab pa’o // slad kyis rgya gar gyi mkhan po ka na ka wa rma daṅ / bod kyi lo tsā ba pa tshab ñi ma grags kyis rgya dpe gsum la gtugs nas legs par bcos pa’o //）。このような場合、現存のチベット語訳テキストから複数のサンスクリット写本の情報を復元するのは不可能である。

3　文化審議会国語分科会『常用漢字表の字体・字形に関する指針』、平成 28 年 2 月 29 日。http://www.bunka.go.jp/seisaku/bunkashingikai/kokugo/hokoku/pdf/jitai_jikei_shishin.pdf

4　CHISE の Web サイト「文字情報サービス環境 CHISE」（http://www.chise.org/）に関連情報がまとまっているので、参照されたい。
5　Chaon モデルについては、守岡知彦「CHISE のデータ形式 Ver.0.1」（2017 年 8 月 21 日）参照。http://git.chise.org/~tomo/character/chise-format.pdf
6　守岡知彦「CHISE 文字オントロジーのための漢字字体・字形粒度の情報記述に関するガイドライン Ver.0.9.1」（2016 年 3 月 11 日）による。http://www.chise.org/specs/ggg_v0.9.1.pdf
7　前掲「ガイドライン」p.1.
8　Pat Riva, Patrick Le Bœuf, and Maja Žumer, “IFLA Library Reference Model A Conceptual Model for Bibliographic Information,” 2017. https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf
9　国際図書館連盟（IFLA）The International Federation of Library Associations and Institutions. https://www.ifla.org/
10　詳しくは、Work の scope notes を参照（前掲書、p. 21）。
11　木村麻衣子「IFLA LRM の漢籍への適用―― work と aggregate を中心に――」