中国古典文献における画像と電子テキスト処理

関西大学文学部　二階堂善弘

前言

　一般には漢籍と称されることの多い中国の古典文献については、近年膨大な数の電子テキストやデータベースが作成され、研究上不可欠なツールの一つとなっている。

これには、コンピュータやネットワーク技術の発展が大きく寄与している。すなわち文字コードが発展した結果、多くの漢字が使用できるようになり、またインターネットの発展により、世界中の研究機関のデータを扱えるようになったのだ。そして、『四部叢刊』や『四庫全書』などの大規模な叢書のデータベースが販売され、中国学という学問自体が大きく影響を受けるようになった。

ただ、そういったツールが一般化することによって、新しい問題が発生し、またそれに伴う考え方の変化が起こっている。小論では、特に電子テキストと画像データの問題に着目し、漢籍のデータ化について些か検討を加えたい。

1．漢籍データと文字コード

中国の古典文献を電子テキスト化、あるいはデータベース化する場合、よく問題とされたのが文字コードであった。

当初パーソナルコンピュータ上で一般に使われたのは、日本ではJISX 0208であり、中国ではGB2312であった。これらはいずれも6千字程度の漢字が使用でき、日常的な語彙を表記するには十分であったが、古典文献を記述するには不十分であった。台湾・香港地区ではBig5が使われ、これは1万3千程度の漢字を使用することが可能であった。そのため、多くの電子テキストやデータベースは、Big5に基づいて作成された。例えば、台湾中央研究院の「漢籍電子文献」(1)などは、Big5を用いて作成されている。しかし、それでも多くの外字を作成して処理する必要があった。

これらの文字コードは互いに交換することが難しく、併用することも基本的には出来なかった。そのため、データの流通において常に困難がつきまとっていた。例えば、GB コードで電子テキストが提供されていた場合、それを JIS コードに変換し、なおかつ簡体字を日本における新字体に直してから処理する、という必要があり、甚だ面倒であった。

Unicode が制定され、Windows などの OS に実装されることによって、こういった作業は徐々に不必要になっていった。

まず Unicode の UCS-2 は約 2 万字の漢字が使用できたが、これは GB と JIS と Big5 を統合したものであり、実際には多くの漢字が重なることとなった。また古典をデータ化する場合は、結局は Big5 とほぼ同じ字数であったと言える。とはいえ、データの交換は遙かに容易になり、また検索する上での利便性はかなり高まった。

むろん、UCS-2 にはその漢字統合の方法に大いに問題があった。「説（8AAC）」と「說（8AAA）」が別の文字として別のコードを振られたりすることは、検査の効率性から見ても大いに疑問であったし、また「呉（5449）」「  （5433）」「吴（5434）」をやはり別の文字とするのは、一応は理由があるとはいえ、穏当を欠くものであると言える(2)。一方で、「与」や「画」のように、日本と中国で明らかに異なる形を持つものを同一のコードとしてしまうなどの問題もあった(3)。

後に Unicode は拡張されて、全体の領域が約 110 万字になり、多くの漢字が追加されていった。そして使用可能な漢字数は約 9 万字にのぼった。またこの拡張 Unicode は、Windows 以外の OS やアプリケーションにも実装されることになり、事実上文字コードのスタンダードな位置を占めるようになった。この拡張された漢字については、その多くの漢字は異体字である。ただ異体字における相互関連はほとんど考慮されないまま拡張されているところから、依然として多くの問題を残している。

とはいえ、データの交換の容易さなどを考えれば、この拡張 Unicode を使って電子テキストや データベースを作成することが推奨されよう。しかし、使用フォントの問題や、適正な検索ツールを欠くことなど、まだまだ解決すべき点は多い。

---

2. 電子テキストの形式について

Unicode のどのバージョン、すなわち、UCS-2 のみに止まるか、拡張 A までにするか、拡張 B まで含めるか、或いは外字を用意するか、といった問題はあるものの、文字コードに Unicode を使用して漢籍データを作成することについては、それほど異論は無くなりつつある。現在の問題は、その形式をどうするか等の議論に移っていると思われる。

そもそも、これほどに文字コードの議論が高まり、また多くの異体字が拡張漢字 B に追加されたこと自体、「見たまま」のテキストをなるべく反映させたいという意図があった。これであれば、電子テキストを見るだけでそのおおもとの状態が明確になるというわけである。

これについては、幾つかの点で問題がある。まず検索である。現在でも Big5 系のツールで作られた電子テキストは、「說（8AAA）」を使用して作成されているため、JIS コードが中心の IME を使って検索しても、「説（8AAC）」とは異なる文字であると認識され、検索結果に反映されない。これくらいであればまだそれほどの問題ではないが、「 ▪ （279D8）」 の字となると、これを「説」の異体字として検索するには、よほど洗練されたツールが必要になると考えられる。「見たまま」を尊重するのもどこまでやるべきか問題になる。さもなくば、際限なく外字を作成することになろう。むろん、それを承知した上で、敢えて見たままを尊重して作成しようという試みもある。それはそれで一定のポリシーによって作成されるのであれば構わない。

ただ我々が日常的に使う漢籍にも色々な種類がある。我々はいつも印刷された活字本（排印本）をも漢籍として扱っているが、この場合は、異体字の多くは通行の字体に直されてしまっている。また編者によってデータが校正されていることも多々ある。このような活字本のデータと、電子テキストを同列に扱うこともちろん可能である。

かつて筆者は『三国志平話』と『武王伐紂平話』の二種のテキスト（『全相平話』二種）を電子化したが、この時はほぼ活字本を作成するのと同じ要領でテキストを作成した(4)。ただこの場合、句点を施し、人名地名の誤りを正すなど、かなりの労力を必要とした。これについては原本のテキストをかなり改変しているため、原本との相違はかなり大きく、ある意味別の本であるともいえる。そのため、このデータを使うにあたっては、原本の画像との併用が欠かせない。とはいえ、インターネット上において画像を公開するに当たっては、所蔵機関の許可など、クリアしなければならない問題が幾つか存在した(5)。

ただこのように、電子テキストやデータベースの作成者がどのようにデータを改変しているか、画像によって確かめることができれば、電子テキスト自体は、必ずしも原本そのものの反映を必要とはしないわけである。

むろんこのようなデータ形式も多くの点で問題を抱えている。特にこの『全相平話』二種のテキストにおいては、原本の誤り自体が重要な情報を含むこともあり得る。例えば、『三国志平話』においては、「司馬懿」はほぼ「司馬益」と誤って記されるが、この混同は北方中国における入声の消滅現象を反映するものである。そもそも『三国志平話』がどこで書かれたかについて、これは有力な情報を与えるものかもしれない。しかし電子テキストだけを見ている場合は、そもそもその差異についての情報が無いため、この問題が存在するかどうかすら認識されない可能性がある。

このため XML などのマークアップ言語を使って、電子テキスト自体に情報を付加するという考え方もあり、多くのデータが作成されている。これによれば、校正前の原本のデータと、校正後のデータとを両方持っているため、双方の情報を参照することができる。また異体字についても、通行字体と異体字の両方をデータに含めることが可能である。

これは非常に有用な手段であると考えられる。但し、現在のところ、どこまでの原データを持ち、どこまで情報を付加していくのか、データ作成について明確ではない点も多い。固有名詞をどう処理するかについても、様々なパターンが考えられる。いろいろ情報を付加できるだけに、かえってその構成については事前によく考える必要があると思われる。ただ将来的には、画像データに XML などのマークアップ言語を付したデータベースの作成も検討すべきであると考える。

---

3. 明刊『封神演義』データの作成

本研究(6)において、電子テキスト作成の対象となったのは明刊の『封神演義』である。このデータベースを作成することによって、具体的に画像と電子テキストの関連についての検討を行った。『封神演義』は、明版は独立行政法人国立公文書館内閣文庫に蔵するものが唯一のものである。電子テキストは幾つか存在するが、みな流布本である清代の「四雪草堂」系統に基づいたものである。

画像とテキストのデータ処理については、『四庫全書』や『四部叢刊』における方法が優れたものとして評価できる。これは北京の書同文公司によって作成されたものである。

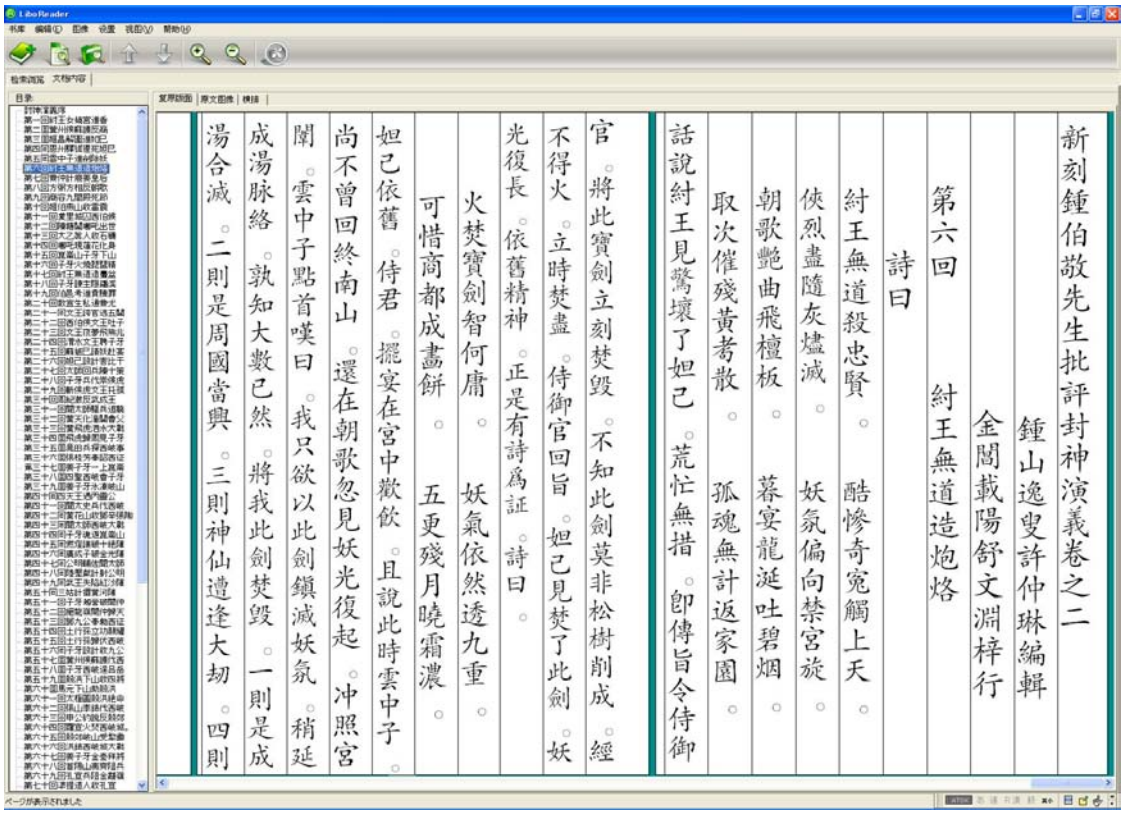
電子テキストは Unicode で作成されており、どの国の OS においても動作可能である。また画像データと電子テキストが同じレイアウトで作成されており、検索したデータをそのまま画像を呼び出して確認できる。画像とテキストデータの有機的な結びつきという点では、非常に洗練された動きを持つ。ただ幾つかのデータは、異体字をやや過剰に処理しており、また検索されたデータの確認については、毎回版本の形式で確認する形となり、逆にそれが面倒に感ずる面もあった。またこれは当然であるが、データとして利用できるのは、あくまでもこれらの叢書類にもともと含まれるものに限られている。

『封神演義』の電子テキスト化に当たって、当初画像データとのリンクについては、インターネット上のみのデータを想定していた。すなわち『全相平話』二種と同様の形である。しかし電子データ作成を『四部叢刊』を担当した書同文公司に依頼したところ、関連企業の創新力博公司の協力により、画像と電子テキスト のリンクを行うプログラムのプロトタイプを使用できることになった。

このプログラムは「青典閲読器」といい、特定の形式で作成したデータを取り込めば、『四部叢刊』で使用している形式でデータを運用できるものである。各研究者がそれぞれ作成したデータを登録することにより、『四部叢刊』と同様の形で検索し、画像とリンクしたデータとして扱うことが可能である。但し、この形式は画像と電子テキストをリンクした独自のものであり、汎用性には乏しい。

この青典閲読器では、版本と同じレイアウトによる電子テキスト閲覧、それにリンクした版本画像閲覧、さらに電子テキストだけの横書き形式による閲覧が可能 である。単に電子テキストとして処理する場合は、この方が簡便性があり、優れていると思われる。但し現状では横書きテキストと画像データのリンクは行われていない。

現在はプロトタイプであるために、検索と閲覧、それにテキストと画像データのリンクの機能が主であるが、将来的にはこれが拡張される可能性が高い。



青典閲読器

今後の発展に期待できるものとしては、「校正」機能がある。これを使用すれば、二種の異なる版本を入力して、その校正を自動的に行うことができる。もっとも、通俗小説データの場合は、段落ごとに大きく内容が入れ替わっている場合があるため、単純に応用するわけにはいかないと思われる。

結語

今後はこのようなツールを使うことにより、画像データと電子テキストの有機的な運用が可能になると思われる。しかしながら、現段階では単なる検索と閲覧のみに止まっており、今後はより深化した使用法を、むしろ人文学に携わる者が積極的に提案して反映させていくことが必要であろう。

なお本研究で作成した電子テキストについては、インターネットによる公開を予定している。

注

1．台湾中央研究院「漢籍電子文献」

2．この問題については、拙論「『三国志平話』データベースの構築－Unicode テキストの利点 と問題点－」（『東洋文化』東京大学東洋文化研究所・第 79 号・1999 年）57～70 頁、及び拙論「『武王伐紂 平話』データベースについて」（『漢字文献情報処理研究』創刊号・2000 年）4～8 頁にて論じた。

3．これについては拙論「多漢字・多言語 Web サイト構築における諸問題」（『漢字文献情報処理研究』第 3 号・2002 年）30～33 頁参照。

4．1999～2000 年 度学術振興会科学研究費補助金・奨励研究 A「全相平話二種データベースの構築」課題番号 117102471（研究代表者・二階堂善弘）

5．当サイトにおける公開。

6．2003～2004 年度文部科学省科学研究費補助金・特定領域研究（2）「東アジアの出版文化」公募研究「中国古典文献における画像及びテキストデータ処理の諸問題」課題番号 15021208（研究代表者・二階堂善弘）