画像とテキストの位置づけ

二階堂　善　弘

Positioning of image data and text data

NIKAIDO Yoshihiro

I have written articles about text processing and image processing. However, recent information technology has improved. Google provides BERT, Baidu provides ERNIE, and the situation of natural language processing has also changed. The situation of image processing has also changed a lot. So we have to fit the current situation.

キーワード：画像データ、テキスト処理、データ処理、文字コード

はじめに

筆者は以前、「中国古典文献における画像と電子テキスト処理」という文章において画像とテキストの問題を論じたことがある。[1] また、別に「電子テキストの発展とさらなる問題点」でその後の問題を論じた。[2] しかしさらに時間が経過し、当時とはインターネットに流れる情報の量も、またデータを扱うコンピュータの性能も全く異なる状況となってしまっている。そもそもテキスト処理自体が、当時とは全く異なる様相を呈している。ここではその変化をふまえ、画像とテキストの処理について、改めて考えてみたい。[3]

1. 文字コードとテキスト

筆者はかつて、ユニコード（Unicode）において、必要以上に異体字を区分しすぎたために、

1）二階堂善弘「中国古典文献における画像と電子テキスト処理」（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 報告書）2004 年 1-10 ページ

2）二階堂善弘「電子テキストの発展とさらなる問題点」（漢字文献情報処理研究会『中国学と情報化』好文出版 2016 年）2-9 ページ

3）本研究は、2017 ～ 2019 年度「私立大学研究ブランディング事業」、及び 2020 ～ 2021 年度「関西大学研究ブランディング事業」によって行われたものである。

かえってテキストデータが作成しにくくなる問題について指摘した。[4)]

> 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）」「吳（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。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。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。

ここで指摘した問題は、まだ解決しているとは言いがたい面もある。データベースやテキストデータを作成する場合、拡張されたエリアの異体字についてはあまり考慮されていない。

むろん、異体字自体に重要な情報が隠されている場合があるし、避諱や欠筆など、年代を測るための情報というのも必要であろう。その点は軽視してはならないと考える。しかし、いままでの古典の活字本を取りあげた場合、異体字はそれほど考慮されていなかったと考える。

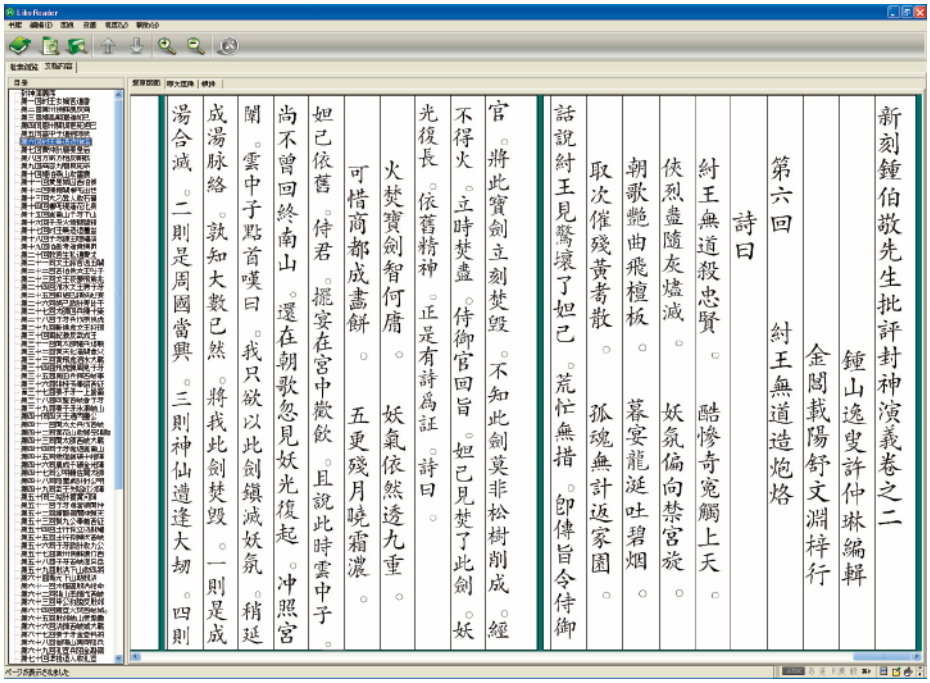
たとえば、中華書局の『二十四史』や、同じく中華書局の『新編諸子集成』などでは、異体字があっても、多くは通行の字体に直してしまっている。学術的にそういった資料を利用してきても、これまでは全く問題はなかった。テキストの検索を考慮するならば、異体字についてはもっと思い切って統合してしまってよいのではないかと考える。

日本社会における漢字処理については、よく「辺」の異体字が問題になるが、「邊」「邉」以外にも多くの異体字をコード化してしまうのはやり過ぎではないかと思う。手書き文字はあくまでも手書きであって、その細かい差異は活字にする時は無視して構わないはずである。手書き文字と、印刷字体の用途を適宜使い分けてこなかったことが、文字コードの使い方にも混乱を与えているように思える。

そのため、資料のもとの情報についてはその資料の画像をそのままアップし、テキストについては、可能な限り通行字体に直したものをアップするというのが、やはり現在でも有効なのではないかと考える。

また「中国古典文献における画像と電子テキスト処理」では、『封神演義』のデータを例に、そのテキストデータと画像の同時利用について提示した。

4）前掲二階堂善弘「中国古典文献における画像と電子テキスト処理」2-3 ページ



青典閲読器による『封神演義』テキストの表示

青典閲読器による『封神演義』の画像データ表示

一方で、異体字が統合されてしまうことについての問題点については、「電子テキストの発展とさらなる問題点」で論じた。これについて、筆者は次のように書いた。[5)]

> 『封神演義』については、基本的に本文は文言調で書かれていたために、方位詞「裡」はあまり多いとはいえないのであるが、幾つかの例を見るに、「裡 2」にあたる「裡」が使用されていることが目に付いた。或いは『封神演義』の成書は明萬暦頃ではないかと推察される。なお、これらの「裡」について、調べた限り電子テキストはほぼすべて「裏」に統一してしまっていた。『三国志平話』などの平話資料では、かなり「裏」で占められていたと記憶する。この場合は、電子テキストとの乖離は少なくてすむかもしれない。「裡（25683）」「裡（88E1）」の相違と、その統計的処理などは、本来はまさに電子テキストが最も威力を発揮する場面であるはずである。しかし残念ながら、現在の「正字化」してしまっている電子テキスト或いはデータベースにおいては、これが不可能となっている。筆者がこれまで提唱してきた「電子テキスト＋画像」のデータベースにおいても、画像でチェックするだけではなく、異体字それぞれのカウントが必要なのであるため、対応が難しいであろう。

異体字が持っている有用な情報が、字体を統一することで失われてしまうということは、確かに起こりうるので、そのマイナス面については確かに考慮しなければならない。これらの問題については、電子テキストの実例を参照しながら、さらに検討を重ねる必要があるだろう。

2. テキスト処理について

とはいえ現在では、さらに電子テキスト処理の考え方が変わっているように思える。ひとつには、自然言語処理の技術が進み、大量データを処理する手法が一般化していることが要因ではないかと考える。

自然言語処理（Natural Language Processing）では、Google 社の提供による BERT（Bidirectional Encoder Representations from Transformers）が登場したことが巨大なインパクトを与えた。[6)] BERT の前にもいくつかの有用な機械読解が試されたが、BERT が出たことですっかり様相が変わってしまった。BERT はテキストの一部をマスクして、穴埋め問題を行うことで言語モデルを学習するものである。いまやギガバイトクラスの巨大なデータを扱って機械処理することは、当たり前のこととなっている。

ただ、BERT の問題としてはアルファベットで記述された言語については強いものの、漢

5）前掲二階堂善弘「電子テキストの発展とさらなる問題点」5-6 ページ
6）西田京介「機械読解による自然言語理解」（情報処理学会『情報処理』Vol. 62 No. 10・2021 年）7-11 ページ

字を使用する言語の単語レベルに関しては弱いところがあるという問題があった。百度（パイドゥ）はその点を克服すべく、新たに中国語向けにチューニングした ERNIE（Enhanced Representation from kNowledge IntEgration）を開発し、提供している。[7] これにより、おそらく漢字文献の機械処理は進むものと予想される。ただ、当面は漢文などの古典中国語より、現代中国語のほうの解析が進むのであろう。

このほかに、UD（Universal Dependencies）による試みがある。UD は多言語横断的なプロジェクトであり、係り受け構造を解析して言語処理を行うものである。この UD の古典中国語を解析するプロジェクトについては、筆者も参加している。[8]

このような新しい自然言語処理の動向を見ると、そこで必要とされているのは、大量の電子テキストデータ、あるいはコーパスであると考える。

現在は、多くのサイトが電子データを公開しており、個人でもギガバイトクラスの電子テキストを収集できるようになっていると考えられる。

このようななかで、電子テキストのあり方も大量での機械処理を前提としたものに変わってくるのであろうと考える。大量の機械処理を前提とするならば、むしろ異体字を整理して通行字体に改めるほうがやりやすいかもしれない。ただ、これについてもまだ検討が必要であろう。

3. 画像処理について

画像処理について、筆者がかつて想定していたのは、版本の見開きデータと、書籍の関連データのみであった。その後、東アジアの寺廟を撮影するに当たって、その寺廟の写真を画像データとして使うことを予定していた。

しかし現在では、大量の映像データがインターネット上を行き来するような状況になっている。画像ファイルの容量を気にする時代でもない。さらに、版本データなどの古典資料についても大量のデータが提供されることになっている。

トリプルアイエフ IIIF（International Image Interoperability Framework）で提供されている仕組みによって、画像のアーカイブズに対する考え方も変わった。[9] 各地の図書館や研究機関のサイトにある画像データを、これまでは各サイトのやり方で検索するしかなかったものが、IIIF の登場によって横断的に利用できることになった。また所蔵機関が異なっていても、

7）机器之心「中文任务全面超越 BERT：百度正式发布 NLP 预训练模型 ERNIE」（https://www.jiqizhixin.com/articles/2019-03-16-3）、この命名は、あるいは「セサミ・ストリート」のマペットの「アーニーとバート」からのものであろうか

8）その詳細については、安岡孝一・Christian Wittern・守岡知彦・池田巧・山崎直樹・二階堂善弘・鈴木慎吾・師茂樹「古典中国語（漢文）の形態素解析」（『東洋学へのコンピュータ利用』27 巻 2016 年）3-14 ページ、および安岡孝一・Christian Wittern・守岡知彦・池田巧・山崎直樹・二階堂善弘・鈴木慎吾・師茂樹「古典中国語（漢文）の形態素解析とその応用」（情報処理学会『情報処理学会論文誌』vol. 59 No. 2・2018 年）323-331 ページなどを参照

9） IIIF については https://iiif.io/about/ などを参照

網羅的に比較することが可能となったのである。これは大きな変化である。

IIIF ビューアによる『老子化胡経』の表示（関西大学所蔵）

このように状況が変化し、かつてインターネットで主流であった JPEG などの圧縮データの画像は、現在でもなお主流でありつつも、より高精度の画像に置き換わりつつある。版本のデータなども、どうにか判別可能な白黒のデータがほとんどであったものが、現在はカラーで高精度のものが提供されるようになった。インターネットで送られる容量が増加したので、TIFF などの高精度ファイルを使ってもそう問題はない。むしろ、高精度ファイルのほうが好ましいということにいまは変化しているようである。このため、画像ファイルのあり方も、「用途に応じて」変化させるべきなのだと考える。

4. 寺廟データについて

筆者は中国や東南アジアの各地で、寺院や廟の撮影を行い、これまで多くの画像のストックを得てきた。その一部分を公開したのが、「寺廟データベース」である。ただ、これはすでに時代遅れのフォーマットとなっている。

おそらく、IIIF ビューアに対応するようなやや高精度の写真データを準備すべきであると考えているが、いまだ十分に対応できていない。

ただ、様々なツールや方法が発展した現在では、新たに寺廟データのあり方を考え直し、より柔軟な設定が必要であると考える。

これまで、寺廟データはある程度の系統だった配置を想定していた。たとえば、東南アジアの廟、台湾の廟といった形で地域に分けて整理する。あるいは、媽祖廟、関帝廟、玄天上帝廟、斉天大聖廟のように、再試する主神ごとに分けて整理する。あるいは、広東系、福建系、潮州系のように、その伝来元を考えて整理する。様々なパターンを考えることが可能であった。ただ、現在ではそういった系統などについては、写真データに付加情報を与えることで、あとはユーザの側でランダムに検索して並べることも可能となっている。それほど系統を意識しなくても、不便にはならないものと思われる。

台南の媽祖廟

マラッカの媽祖廟

長崎崇福寺の媽祖堂

おわりに

以上、現時点でのテキスト処理と画像処理の問題について検討を行った。現状では、まだまだ課題も多い。ただ、これまでの単純なテキスト処理と画像処理の考え方ではもはや通用しないことも認識した。これらへの対応については、今後の課題としたい。