電脳中国語フォーラム
公開シンポジウム

2002 年 6 月 30 日
於早稲田大学文学部

セッション I　**漢字文献情報処理の基礎**

中国学情報化の現状と課題

慶應義塾大学経済学部専任講師
漢字文献情報処理研究会（http://jaet.gr.jp/）副代表

千田　大介

http://wagang.econ.hc.keio.ac.jp/

§1. 熱狂と憎悪の狭間

❶ 人工知能の夢

1950・60 年代：鉄腕アトム・HAL ＝人工知能の夢
1970 ～ 90 年代：人工知能研究の頓挫：**夢は幻想に** ⇨ Ⓐ

- 完全無欠な人間の比喩としてのコンピュータ
  ➡ コンピュータへの熱狂と妄信
- 人間を代替する存在としてのコンピュータ
  ➡ コンピュータへの憎悪と無視

いずれも誤り：**適正な情報化の障害** ⇨ Ⓑ

❷ コンピュータの適性

コンピュータは人工知能たりえない＝コトバが苦手

- 数値化できないコンテクスト
- 身体の欠如
- 脳の膨大な情報量

機械にできること＝人間の業績にならない
➡ コンピュータは研究者・教育者の仕事を減らさない

∴コンピュータが人文学・語学の教育・研究を根底から覆すことはあり得ない

- コンピュータ＝計算機
- コンピュータで文字は数値として扱われる
  ➡ **文字処理ツールとしての適性**
- インターネット：巨大な情報交換装置
  ➡ **情報ツールとしての適性**

∴コンピュータのツールとしての適性を見極めた使いこなしが必要：コンピュータの使用を前提にしない ⇨ Ⓒ

§2. 人文学情報処理の三本柱

❶ 研究対象のデジタル化
❷ 研究方法・ツールのデジタル化
❸ 研究成果のデジタル化
⬅ インターネットを通じた情報流通

Ⓐ 人工知能の幻想

1972 年に出版された『**コンピュータには何ができないか**』（ヒューバート L. ドレイファス著。黒崎政男・村若 修訳、産業図書、1992 年）は、人工知能批判の古典的名著。『**メディアラボ**』（スチュアート・ブランド著、室謙二・麻生九美訳、福武書店、1988 年）：

「人工知能はどうなったんでしょうか？」とわたしはたずねた。「開発研究が始まって 30 年たって、まだ人工知能はなんとかして成功しようとがんばっているのですね？」「人工知能はつねに遠ざかっていく目標として設定されたのだ」とミンスキーは答えた。

マービン・ミンスキーは、著名な人工知能開発の先駆的研究者。

Ⓑ 適正な情報化の障害

アメリカにおいて、コンピュータテクノロジーへの妄信が教育に与えた悪影響については、『**コンピュータが子供をダメにする**』（クリフォード・ストール著、倉骨彰訳、草思社、2001 年）に詳しい。わが国で現在進められている教育情報化計画の中には、ストールの批判する失敗例と驚くほど似通ったものが、多々見受けられる。

Ⓒ コンピュータと生産性

Robert J. Gordon. “Has the “New Economy” Rendered the Productivity Slowdown Obsolete?” 1999(http://faculty-web.at.nwu.edu/economics/gordon/334.html) によれば、アメリカの産業の大半で情報化による効率低下が発生している。語学教育でも、10 分のドリルソフトの作成に 1 時間かかり、遠隔授業のために TA を雇用するのであれば、そのマンパワーを個別指導に振り向けた場合との教育効率比較が欠かせない。

❶ 研究対象のデジタル化

- 20c 末：台湾中央研究院：学術データベースの草分け
- 21c：中国の電子テキスト構築：技術力・労働力・文化部の指導：大規模データベースが続々と作られる　⇨ Ⓓ
  『四庫全書』：7 億字　『四部叢刊』：1 億字
  中国基本古籍庫：20 億字
- 電子テキスト構築：中国への委託入力：国際分業
  ➡ **学術情報化への意識と資金力の問題**
- 著作権問題：他者の権利の尊重　⇨ Ⓔ
- 品質をいかに保証するか

❷ 研究方法・ツールのデジタル化

- 一般化された検索機能：自由度が低い
  ➡ コンピュータのスキルが引き出せる情報を左右
- 検索・分析ソフトウエアと研究手法の開発　⇨ Ⓕ
- 辞書データ：同義語・音韻・用語集 etc.

❸ 研究成果のデジタル化

- 電子ジャーナル：論文の題目・全文データベース
  米中台＝利用が常識：電子ジャーナル構築の制度化
  日本＝大幅な立ち後れ：制度化されていない　⇨ Ⓖ
  ➡ **ネット上に存在しない研究は無視される**

❹ インターネットと学術情報発信

- インターネットは基礎教育に向かない：情報を取捨選択する能力が必要：レポートのネット盗作問題　⇨ Ⓗ
  ➡ 高度な学術情報の交換に適性
- インターネットは本来学術情報交換の枠組み

§3. 情報化後進国からの脱却へ

- 情報化：学術水準向上＝評価のためのツール
- インターネット時代：“鎖国”は成立しえない
  ➡ **学術情報化への対応は不可避**
  - 個人：自らが必要とするデータの構築・公開
  - 学会：規格・型式の標準化、学術分野の研究情報集積
  - 国立情報学研究所（NII。http://ge.nii.ac.jp/）
    ：国家的技術サポート態勢　➡ **意識の問題**

∴研究者＝情報発信者としての自覚：適正な情報化の推進

Ⓓ 中国の大規模データベース

以下の大規模データベースは、書同文公司（http://www.unihan.com.cn/）の文献デジタル化技術の成果。詳細は、拙著「中国における古典文献データベースの構築－書同文公司へのインタビューを通じて」（『漢字文献情報処理研究』第二号、2001）参照。また、画像による電子図書館である超星数字図書館（http://www.ssreader.com.cn/）では、40 万冊もの書籍を公開する。

Ⓔ 著作権問題

インターネットを流通する中文電子テキストの大半は、違法データ。

- 著作権保護期間：没後 50 年
- 校訂・注釈：二次著作権が生じる
- 翻刻・影印：所蔵者の財産権

資料編『中文電子テキストの十年』参照。著作権については『インターネット時代の著作権』（半田正夫著、丸善ライブラリー、2001）参照。

Ⓕ デジタル研究手法

⇨ 「N-gram 特集」（『漢字文献情報処理』第二号）など。

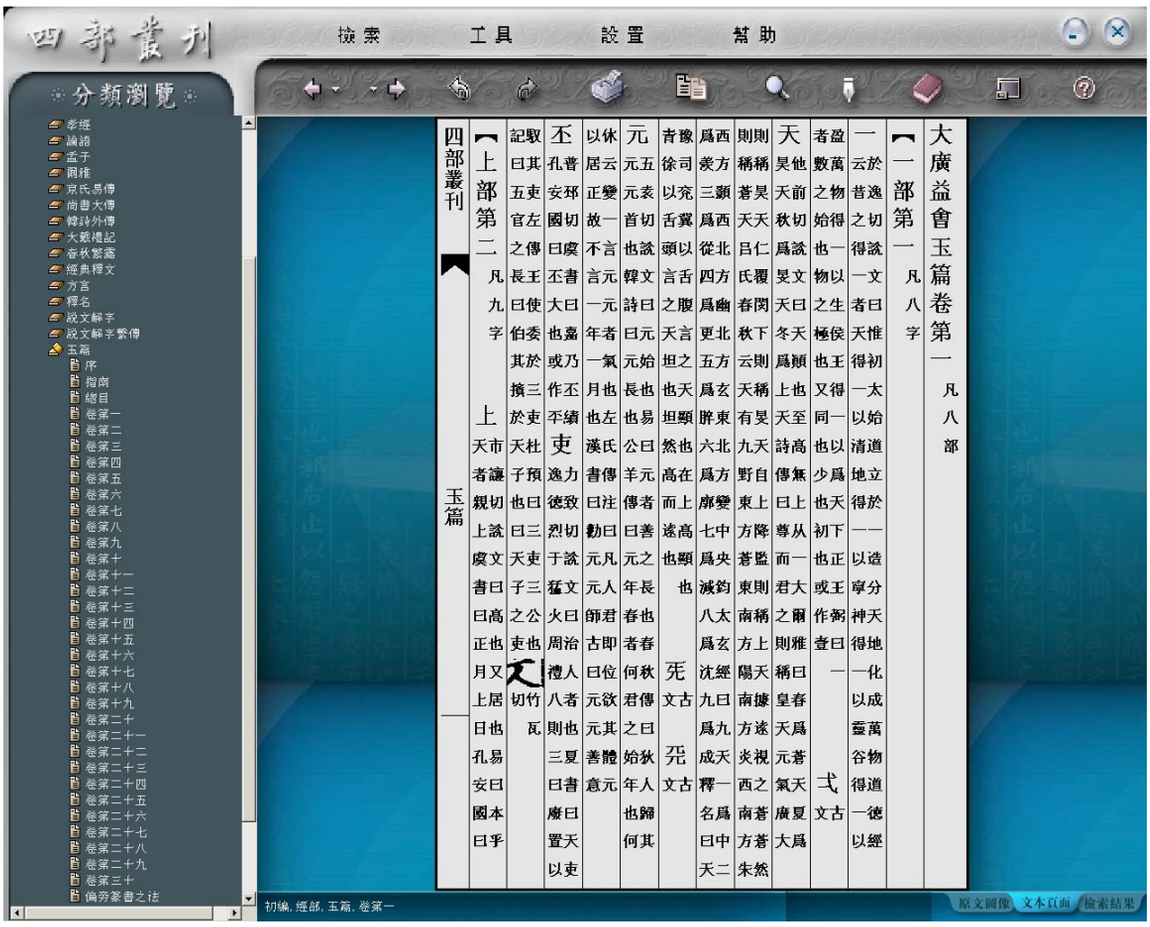
Ⓖ 電子ジャーナル

台湾国家図書館遠距離図書服務系統（http://www.read.com.tw/）は、1991 年以降の台湾・香港の論文のキーワード検索、ダウンロード（有償。作者が許可したもののみ）サービスを提供。中国期刊網（http://www.cnki.net/）では、1996 年以降の中国の論文の提要全文検索、ダウンロード（サイトライセンス）が可能。しかし、上記データベースと契約している日本の大学はごく少数。

Ⓗ インターネットと基礎教育

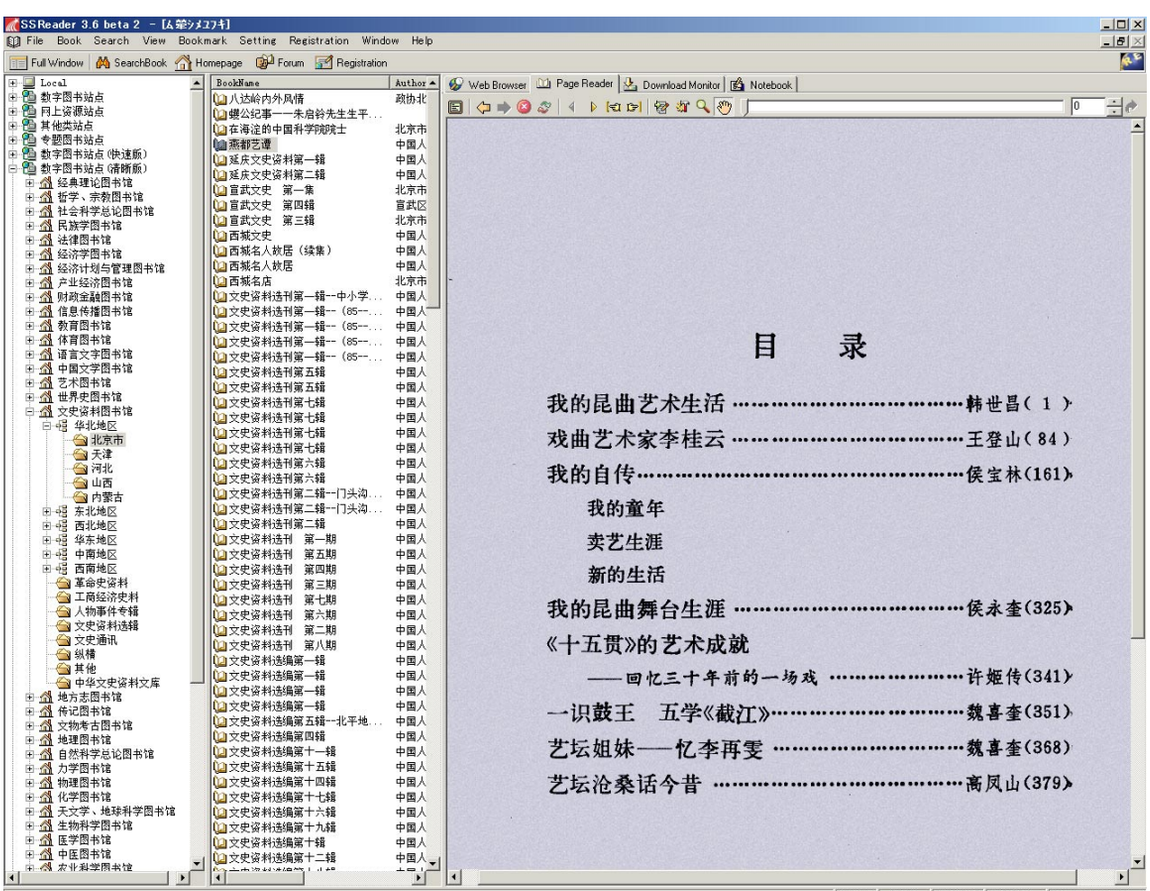
前掲『コンピュータが子供をダメにする』は、インターネットによる学習や遠隔学習の問題点を実例を挙げて詳細に論じた上で、断ずる。

> 遠隔学習とは、三流の教育を受ける（享受する）卓越した手段だ。



⇧ 『四部叢刊』原文及全文検索版

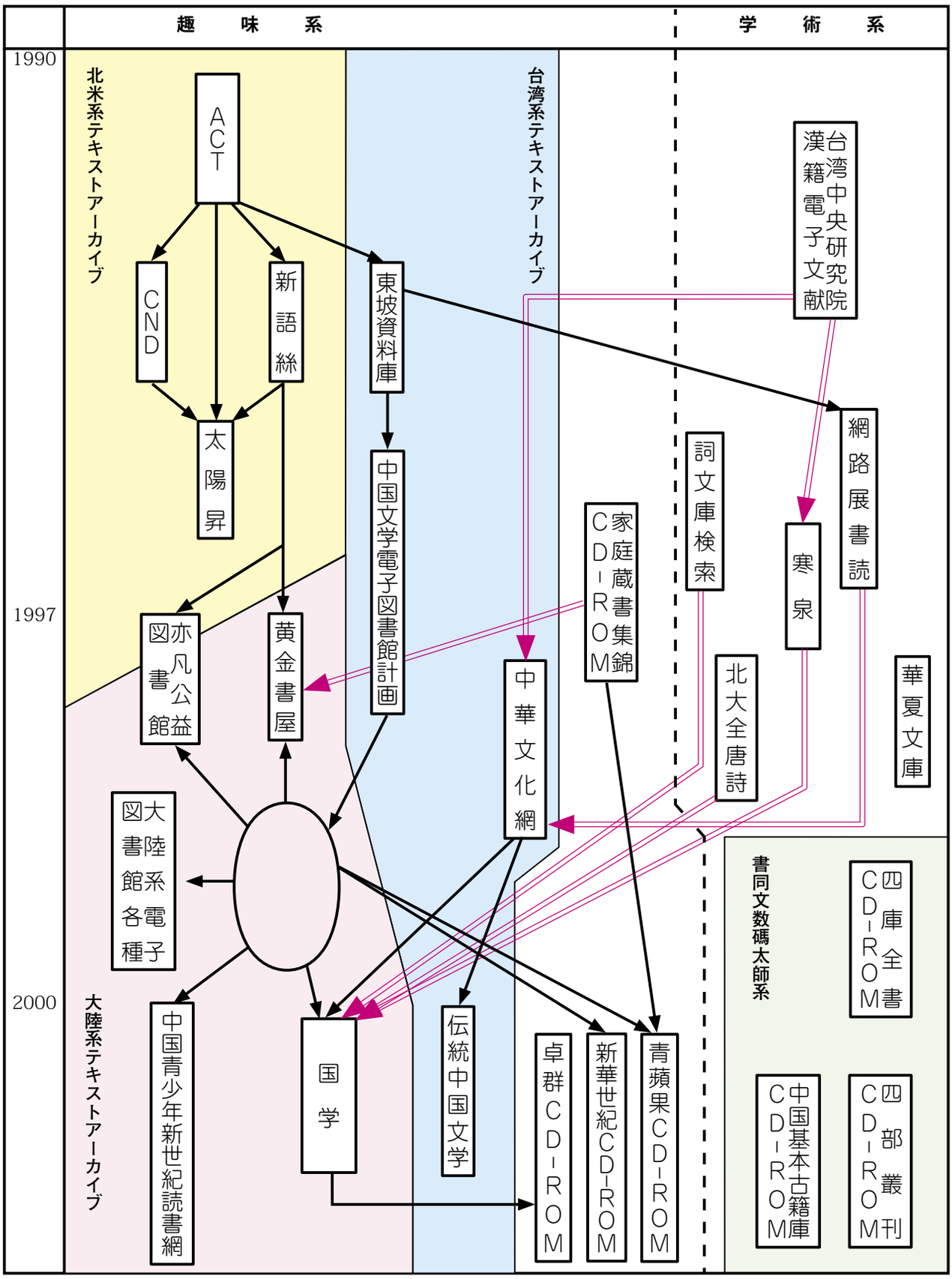
⇩ 超星数字図書館



中国古典文献電子テキスト流転略図

転載　違法性の高い転載

※いずれも推定を含む　※電子テキスト化そのものの違法性は反映していない



中文電子テキストの十年

千田　大介

1. 中文電子テキストの黎明

はじめに

中国古典文献の二十一世紀は電子テキストのデフレとともに幕を開けた。中国のちょっと大きな書店に行けば、二十五史の CD-ROM はわずか三千円、『全唐詩』に到っては五百円で手に入る。CD-ROM だけではない。国学や亦凡公益図書館などのオンラインテキストアーカイブを訪問すれば、さらに多くの古典文献電子テキストを無償で入手できる。五年たらず前、四書五経の電子テキストの入手すらもおぼつかなかった頃とは隔世の感がある。研究者しか読まないようなお堅い古典文献の大全集が、今や、だれにでも手軽に入手できる通俗的商品になってしまったのだ。

読書に研究に、とりあえず手軽に利用できるオンライン電子テキストは、既に十年にもおよぶ発展の歴史をもっており、それは台湾中央研究院・寒泉・香港中文大・四庫全書などの学術データベースの発展史と表裏をなしている。小稿では、この十年間の中国語古典文献電子テキスト発展の経過を、ごくおおざっぱになぞってみたい。もちろん、インターネットを行き交ったさまざまな電子テキストについて完全に把握することは筆者の能力では到底不可能なので、誤りは免れえないものと覚悟している。諸賢の指正を乞う次第である。

東坡資料庫と CND・新語絲
～電子テキストアーカイブの成立

日本でようやくインターネットが普及しはじめた 1995 ～ 97 年頃、中国はまだインターネットに公式接続しておらず、中国語ネットワーカーは欧米や日本などの留学生・華僑と中国プロパー、それと台湾・香港の人びとだけだった。このため、文献入力の担い手はごく少数のボランティア愛好者だけに限られ、電子テキスト構築の歩みは遅々としていた。

この時期、しばしば利用された中国語オンラインテキストアーカイブは

・東坡資料庫 gopher://dongpo.math.ncu.edu.tw/
・CND http://www.cnd.org/
・新語絲 http://www.xys.org/

現在でもサービスは続いているが、中国大陸のテキストデータデフレのおかげですっかり影が薄くなってしまった。

東坡資料庫は台湾中央大学の單維彰氏の提供で、Big5 コード繁体字。CND は北米の民主派よりニュースダイジェスト「華夏文摘」の発行元で、HZ コード簡体字。これらが 90 年代前半にサービスを提供していたのに対して、新語絲サイト（GB コード簡体字）の開設は 1996 年とやや遅れる。北米で大陸より現実路線の同名オンラインマガジンを発行しており、CND とは浅からぬ因縁があり水と油の仲だ。

この三カ所に収録されているデータは、四書五経に『唐詩三百種』『紅楼夢』、ほかは名作古典詩文と古典小説がちらほら、といった具合でおおむね一致する。例えば、四書五経の大半は、John H. Jenkins（井作恒）の入力。オリジナルは Big5 コード繁体字版で 1992-3 年の日付が見える。『孫子』『鬼谷子』などは、カリフォルニア大の張家傑氏の入力。入力は 1991 年に完了している。1990 年代初頭といえば、日本ではパソコン通信が普及しはじめた時代、中国語のバーチャルコミュニティはまだ成立していなかったから、それらはおそらくスタンドアロン環境で入力されたのだろう。

ところで、世界で最初に成立した中国語バーチャルコミュニティーは、1992 年、アメリカインディアナ大学に設置されたニューズグループ、Alt.Chinese.Text（ACT）であるとされる。それまでにも英語のコミュニティーはあったが、簡体字 GB コードを細工した HZ コードによって ASCII コードのシステム上でも中国語情報を交換できるようにした点が画期的だった。欧米の中国人留学生の生活情報交換にはじまり、時事批評、文芸批評、さらには文学作品の発表の場としても活用されるようになり、中国語総合バーチャルコミュニティーとして 90 年代半ばに最盛期を迎えた。

参考
・方舟子「ＡＣＴ的興起」
http://www.xys.org/xys/netters/Fang-Zhouzi/Net/act1.txt

例えば新語絲の『莊子』テキストに Alt.Chinese.Text のヘッダが残り、データの入力・校訂者に ACT で名を馳せた「網文八大家」の面々が見えるように、ACT は電子テキストの入力・交換の場としても機能していたようだ。それらを総合したのが、これらのオンラインテキストアーカイブということだろう。（筆者は当時 ACT を講読していなかったので、このあたりの事情に詳しい方

がいたら、具体的状況をご教示いただきたい。）

新語絲から大陸系テキストアーカイブへ

1997 ～ 98 年には、東坡資料庫と CND は電子テキストの更新をほぼ停止したが、新語絲だけは次々と新たらしい古典文献のテキストデータを公開していった。『全唐詩』を切り取った李白・杜甫をはじめとする唐代詩人の別集、詞のアンロジー、さらには『三国演義』全文のテキストデータ化を完了させるなど古典白話小説のテキストデータ構築もすすめた。四書五経や諸子などは学術利用も考慮して、校訂もそれなりにきちんとなされているようだ。また、著者サイドの抗議を受けて金庸全集を削除するなど、中文電子テキストをめぐる著作権問題にはじめて直面したのも、新語絲だ。

中国では 1997 年のインターネット接続の後、1999 年前後に多くのオンラインテキストアーカイブが出現するが、新語絲のデータはそれらの基本コンテンツとして利用された。つまり新語絲は、北米を中心とした先駆的なテキストデータ入力の成果を集大成して発展させ、現在の中国大陸系オンラインテキストアーカイブへと継承する媒介の作用を果たしたと言える。このことは、中文電子テキスト発展の歴史において、高く評価されよう。

かくて、1999 年には中文テキストデータの構築の主役の座は、北米系サイトから大陸系サイトへと交代し、オンラインテキストアーカイブの仁義無き拡張の時代が幕を開けるのである。

2. 中文電子図書館サイトの勃興

中国のインターネット接続

中国のインターネット接続の歴史は、1988 年にさかのぼる。接続当初は大学間ネットワークだけに限定されたが、1990 年代半ばにインターネットの商用利用が世界的に大ブレイクすると、情報管制に熱心な中国当局といえどもその存在を無視できなくなった。かくて 1996 年、中国国内の基幹ネットワークがインターネットに接続、1997 年には CNNIC（http://www.cnnic.com.cn/）が設立され、中国は正式にインターネットに接続した。

前回にも書いたように、黎明期の中国語インターネット世界の担い手は､在米華人と台湾とであった。しかし、中国が正式にインターネットに接続すると、既存のサイトは中国で爆発的に増えてゆくネットユーザーを無視し出来なくなったし、また、中国国内からの Web を通じた情報発信も次第に増えていった。

その具体的な動きの一つが、中文電子テキストを集積したオンラインテキストアーカイブ、すなわち中文インターネット世界で「電子図書館」と呼ばれるサイトの隆盛である。

亦凡書庫と黄金書屋

この時期に生まれた代表的な中文電子図書館に、亦凡書庫と黄金書屋がある。いずれもサービス開始時期は 1997 年頃。亦凡書庫の本拠はニューヨーク、黄金書屋は中国国内（おそらく上海）で運営されていた。1999 年頃まで、両者は同じような発展の軌跡をたどった。

両者はともに、電子図書館サイトを開設するにあたって、新語絲（http://www.xys.org/）などの中文インターネットの草分け的サイトが公開していた、ネットワーカーがボランティアで入力した古典や近代文学、武侠小説などのテキストデータを借用した。その意味で電子図書館サイトは、黎明期の中文インターネット文化の継承者であると言える。

しかし、それらの新興電子図書館は、新語絲などの従来のサイトとは決定的に異なっていた。電子図書館サイトは学術性の薄い通俗的なサービスで、サイト間の競争も激しかった。このため、各サイトが競い合うように、次々と新たなデータを入力、公開していったのである。

そのスピードを支えたのは、膨大な中国大陸ネットワーカーである。亦凡書庫や黄金書屋は、閲覧者に電子テキストの提供を（あるときは有償で）呼びかけ、それに呼応した中国のネットワーカーは、非常に高価な中文 OCR ソフト（おそらくは海賊版）で、気に入った文学作品を次々と電子テキスト化し、公開していった。そのペースは、中国大陸のネットワーカーの増加と比例するかのように速まっていき、1999 年頃には中国文学作品の電子テキスト構築の主役は、完全にこれらの新興中文電子図書館に移った。

しかし、中文電子図書館は通俗的なサービスであるが故に､娯楽としての読み物の提供に主眼が置かれており、電子化される作品は明清の通俗小説や近現代小説・武侠小説など、読みやすく娯楽性が高いものが圧倒的多数であった。その上、校正は不十分、底本も明示されていないなど、データの質も低かった。

このため、黎明期の比較的少数の意識の高いネットワーカーのコミュニティによって支えられていた時代には確保されていた、版本表示・校正などの電子テキストのクオリティ（学術的利用を考えれば不十分なレベルではあったが)すらも保たれなくなった。また､一部のネットワーカーの研究的興味から入力されていた経書･史書･古典詩文などは、電子テキスト化の対象からはずれてし

まった。

それだけではない。黄金書屋には、古今の哲学・文学作品を PDF 型式で収録した CD-ROM『家庭蔵書集錦』から転用したと思われるデータさえもが公開されていた。また、近現代文学作品には明らかに著作権が生きているものも数多く見られる。例えば、金庸の武侠小説は、新語絲は 1998 年に著者サイドの抗議を受けて全てのデータを削除したが、他の中文電子図書館は堂々と公開を続けている。日本の赤川次郎や田中芳樹などの中国語訳さえも公開している。古典文学にしても、排印本の場合は校点者の二次著作権があるはずだが、もちろん無視である。おそらく、二次著作権の存在すらも知らないのであろう。つまり、矢継ぎ早に電子テキストを公開していくことができた理由の一端は、著作権法の無視にあるのだ。

この問題は、実は黎明期の電子テキストにも当てはまり、現在に至るまで全く解決されていない。相変わらず、ネット上で入手できるテキストデータは、著作権者の許可を得ていない違法データばかりである。このため、たとえ学術目的であろうとも、筆者は中文電子図書館サイトのテキストデータの利用は推奨しない。

ともあれ、公開されるテキストデータの圧倒的な量、そして、中国人ネットワーカーの著作権問題への無頓着は、質の問題を覆い隠し、それらのサイトのアクセス数を増やしていき、ついに電子図書館サービスは中国語サイトの定番として定着することになった。

雨後の筍

ひとたび、亦凡書庫や黄金書屋が成功を収めると、まさしく雨後の筍のように、中国各地にコピー電子図書館サイトが陸続と誕生した。そこには、中国ならではの事情があった。

広大な中国では、国土の隅々まで高速回線網を敷設するのが非常に難しい。現在でも、中国に渡航してインターネットに接続してみると、回線速度の遅さに悩まされるものだ。このようなインフラの遅れは、インターネットとは言っても、快適につながるのは地域のローカルネットワークの中だけ、という現象をもたらした。つまり、人気サービスのコピーコンテンツを各地に置くことは、地域ネットワーカーのニーズに合致するのだ。

とりわけ電子テキストはコピーも加工も非常に簡単であるので、データをコピーして再配置するだけで、お手軽に電子図書館サイトを立ち上げることができる。

こうして、1998 ～ 99 年には、中文電子図書館サイトは乱立の時代に突入した。大唐中文、書路、書香門弟、OCR 書城・・・その数は、50 を超えた（詳しくは Kanhoo! Web Srarch の「オンライン書庫、テキストデータ」カテゴリ http://jaet.gr.jp/kanhoo/web/home_10_03_20.html を参照されたい。そこに登録されているサイトも、あまたの電子図書館サイトの一部に過ぎない）。だが、独自にテキストデータを入力・公開しているものは、おそらくその四分の一にも満たない。大半のサイトは、他のサイトからのコピーデータだけで成り立っている。

面白いことに、亦凡書庫にしろ黄金書屋にしろ、掲載データのコピーは禁じていなかった。自らもそもそもが新語絲のコピーから出発したのだけに、禁止には意味のないことを充分にわかっていたのだろう。その代わり、テキストデータ末尾の入力者のクレジットは保存する、というルールが自然発生的に定着した。このルールは現在も継承されている。現在の電子図書館サイトの中には、クレジットの無い電子テキストもあるが、それらの大半は黎明期に入力されたデータ、あるいは台湾系のデータであり、出自が違うものである。

亦凡書庫と黄金書屋のその後

亦凡書庫は 1999 年に亦凡娯楽信息網の一コンテンツとして再編され、名称も亦凡公益図書館（http://www.shuku.net/）に改められている。相変わらず、古典から近現代まで数多くの電子テキストを入力・公開しているが、近頃は所謂オンライン文学作品の募集にも力を入れており、オンライン総合文芸サイトへの脱皮をはかっている。

一方の黄金書屋は、2000 年に上海の ICP「多来米（ドレミ）中文網」にコンテンツを全て売却し、独立サイトとしての活動を停止した。ドレミ黄金書屋からは、著作権的に問題のある近現代文学作品のテキストデータが全て削除され、近現代文学の書評と古典文学テキストデータとにコンテンツが再編された。翌 2001 年、ドレミ中文網は中国での地歩固めを目指す Lycos によって買収された。Lycos 中国の文学（http://wenxue.lycos.com.cn/）が、さまざまな流転をへた末の黄金書屋の姿である。

3. テキストデータロンダリング

学術データと違法コピー

学術利用を念頭に構築された中国古典文献データベースといえば、台湾中央研究院漢籍電子文献（http://www.sinica.edu.tw/ftms-bin/ftmsw3）に尽きる。構築の開始は 1980 年代末にまでさかのぼり、1997 年には、

二十五史や諸子などの古典文献の検索サービスをインターネットに無料開放し、今に至るまで、中国学古典籍データベースの中核的地位を占めている。

しかし、中研院のデータベースは、本来無償ではなかった。大学・研究機関はデータを購入し独自サーバを構築しなくてはならなかったし、個人も中研院に有償のユーザー登録をしなくてはならなかった。その後、中研院はデータベースの無償提供に方針を切り替えたが、それでもまだ有償公開部分が残っているのは、このような経緯から全面無償化が困難であることに起因する。税金を投入して構築しながら高額な使用料を取るデータベースといえば、読者諸氏にも思い当たる事例は多々あろう。我が国ならば泣き寝入りのケースであるが、台湾では違った。中研院のデータを一頁ずつコピーして集積し公開するという、過激な対抗措置に踏み切るユーザーが現れたのだ。

そのような違法データを公開したサイトに、寒泉（http://libnt.npm.gov.tw/s25/）と 中華文化網（http://www.geocities.com/Area51/Hollow/3198/）がある。

寒泉は、十三経や『紅楼夢』の電子化をいち早く進めた陳郁夫氏が構築するれっきとした学術サイトである。『資治通鑑』などの学術レベルをクリアした独自データも多いのだが、同サイトの二十五史データは、中研院からの違法コピーであるとされる。近頃、二十五史の公開を中止したのには、このあたりの事情も絡んでいるのであろう。

中華文化網は、トロント在住の呉恆昇氏が構築したサイトで、有償公開部分を含む中研院のほとんどの文献データを違法コピーして無償で公開している。このため、有償部分の契約が不可能な日本のユーザーの一部には歓迎されたが、しかし、中研院の外字を適当な異体字に置き換えるなど、テキストの精度は落ちる。

二次著作権への不理解

1998 ～ 2000 年頃には、中研院に続いていくつかの学術古典データベースが公開された。香港中文大の華夏文庫（http://www.chant.org/）、台湾元智工学院の網路展書読（http://cls.hs.yzu.edu.tw/）などがそれである。インターネット上に利用しやすい学術データが増えるのは嬉しいが、しかし、それらのサイトも違法コピー問題に頭を痛めることとなった。例えば、網路展書読はサイトにたびたび違法コピーへの警告を掲載している。

このような違法コピーの背景には、著作権への理解不足がある。例えば、二十五史そのものは、著作権が切れている。しかし、現在ネットに流通しているデータは、台湾中研院が北京中華書局の校点本に基づいて入力したもののコピーである。中華書局本は、独自の校訂と評点とを付したテキストであるので、その作業に対して二次著作権が発生している。従って、校点者の死後 50 年後まで、中華書局本を電子化し勝手に公開することはできないのである。

台湾中研院は、当初、台湾で出版された中華書局本のコピー本に基づいて許諾を得ずに入力していたが、その後、中華書局のクレームを受けてライセンス料を支払ったという。しかし、そのデータをコピー・公開したサイトはそのような手続きを一切経ていない。そもそも、底本の表示すらない。

台湾・中国では、二次著作権への理解が非常に遅れている。一次著作権すらもようやく認知され始めた段階であるから仕方のない面もあるが、しかし、違法は違法である。しかも、それらのデータは違法であるが故に、テキストの由来に関する注記が全く無く、品質を保証する主体もない。このため、学術論文に引用するのははばかられるし、また、学術利用できるだけの水準もクリアしていない。

違法データのロンダリング

ひとたび中華文化網などに流出したデータは、ねずみ算式に、子データ・孫データを増やしていった。亦凡公益図書館などの中文電子図書館サイトは、軒並み二十五史のデータを掲載するようになった。大学・研究機関サイトにも、それらのデータを再配布しているところがある。

学術目的を前面に打ち出した電子図書館サイトである国学（http://www.guoxue.com）も、二十五史・『資治通鑑』・『全宋詞』などのデータは、中華文化網や寒泉・南京師範大学などからコピーしたものであると思われる。また、同サイトが中心となって構築をすすめている各種文献データも、二次著作権問題をクリアしているのか甚だ疑わしく、かつ底本の表示もないので、決して信頼するに足るデータであるとは言えない。

また、中国では 2000 年頃から PDF 型式の古典文献 CD-ROM が多数発売されており、例えば二十五史も、複数の出版社から出ている。あの大部の二十五史を、いくつかの出版社が全文入力したはずはなく、いずれも中研院系違法コピーテキストの孫データであると思われる。

国学にせよ、それらの CD-ROM の出版元にせよ、自らの行為が違法であるとの認識はないようだ。前にも書いた二次著作権への不理解とともに、中研院－中華文化網－電子図書館サイトとコピーを重ねるうちに、データ

の出所が定かでなくなり、違法感がなくなってしまったことも、その原因の一つであろう。あたかも、マネー・ロンダリングのように。

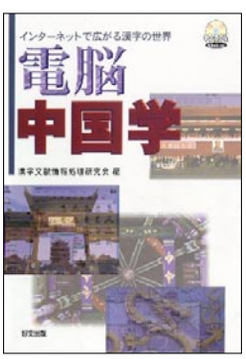
欠かせない品質保証システム

台湾中研院のデータベースでは、複数のキーワードを指定した検索が可能である。しかし、正規表現による複雑な条件を設定した grep 検索や N-gram などの手法によるテキスト解析ということになると、やはりどうしてもテキストデータが必要になる。しかし、前述のように現状では、フリーで利用できる信頼性の高い電子テキストは非常に少ない。従って、今後は、個人や学会・研究機関などが、積極的に著作権問題をクリアした校訂のしっかりしたデータを提供し、かつ責任の所在を明確にして随時データを更新・改訂していくシステムを作り上げる必要があろう。

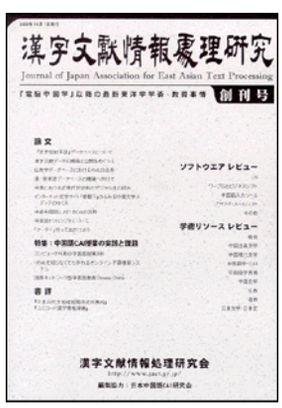
テキストの電子化そのものは、書同文公司（http://www.unihan.com.cn/）などの技術力に定評のある企業に委託すれば、安価かつスピーディーに完成することができる（詳細は、『漢字文献情報処理研究』第二号掲載の拙文をご覧頂きたい。http://www.jaet.gr.jp/jj/2.html）。要は、電子テキストの学術利用に対する意識と、資金力の問題なのである。

※本稿は、漢字文献情報処理研究会メールマガジン第一号・第九号・第十一号に掲載したコラムをまとめたものである。

漢字文献情報処理研究会の刊行物



好文出版　1998 年
A5・300 頁
CD-ROM 付き
定価：2,850 円＋税
ISBN4-87220-023-3



好文出版　2000 年
編集協力：日本中国語 CAI 研究会
B5・152 頁
定価：1,800 ＋税
ISBN4-87220-045-4

好文出版　2000 年
A5・286 頁
CD-ROM 付き
定価：3,200 円＋税
ISBN4-87220-041-1

好文出版　2001 年
A5・368 頁
CD-ROM 付き
定価：3,200 円＋税
ISBN 4-87220-052-7

好文出版　2001 年
編集協力：日本中国語 CAI 研究会
B5・216 頁
定価：2,200 円＋税
ISBN4-87220-051-9