

拓本文字データベースの現状と課題

安岡孝一^{1,a)}

概要：2004年8月に開発を始めた拓本文字データベースは、8年の開発期間の後、漢代～中華民国初期の拓本から180万文字を切り出した、巨大な文字画像データベースへと発展した。しかし、巨大化した結果、さまざまな問題も発生している。本報告では、その現状と今後の展望について述べる。

Character Database of Digital Rubbings: Its Progress and Problems

KOICHI YASUOKA^{1,a)}

Abstract: The author has been developing Character Database of Digital Rubbings since August 2004. Now it consists of 5,000 rubbings and includes 1,800,000 character images from Han Dynasty to Xinhai Revolution. Such gigantic image database has its own difficulty to manage. In this report the author reveals the problems on the database and shows the way to control it.

1. はじめに

拓本文字データベース^{*1}は、京都大学人文科学研究所所蔵の漢代～中華民国初期の文字拓本5,000枚から、およそ180万文字の画像を切り出し、釈文としての検索を可能にした漢字画像データベースである。2004年8月のデータベース開発開始から、2005年2月のWWW公開を経て、2011年3月までに文字拓本4,750枚の文字切り出しを完了した。現在は、開成石経拓本250枚のデジタル画像化と、文字切り出しの作業中である。本稿では、拓本文字データベースの開発過程および現状と、拓本文字データベースが抱える課題および今後の展望について述べる。

2. 拓本文字データベースの開発過程

2.1 開発開始に至るまで

京都大学人文科学研究所は、現在、約10,000点の拓本資料を所蔵している。これらの拓本は、必ずしも美拓では

ないものの、漢から清に至る主要な石刻をほぼ網羅しており、日本国内でも第一級の資料だと考えられる。これらの拓本のデジタル化を計画するにあたり、拓本のデータベースに対してどのような要件があるのかを調査したところ、おおまかに以下の3つが挙げられた。

- (a) 拓本の画像データベース
- (b) 釈文の全文データベース
- (c) 漢～清にまたがる文字データベース

(a)の要件は、古典籍の図書館である人文科学研究所附属東アジア人文情報学研究センター(当時は附属漢字情報研究センター)では、ごく自然な要求だった。(b)の要件は、むしろ中国学の研究者からの要求であり、その時代の資料に何が書かれているかを検索したい、ということである。(c)の要件は、漢字オタクである筆者の願望である。漢字オタクは、書かれている内容よりも、むしろ漢字の字体そのものが見たいのである。

しかし、3種類ものデータベースを構築する資金もパワーもない。そこで筆者は、これら(a)(b)(c)を別々のデータベースで実現するのではなく、1つのデータベースでこれら3つの要件を満たすことにし、拓本文字データベースを構築することにした。

¹ 京都大学人文科学研究所附属東アジア人文情報学研究センター
Institute for Research in Humanities, Kyoto University,
Kyoto 606-8265 JAPAN

a) yasuoka@kanji.zinbun.kyoto-u.ac.jp

^{*1} 拓本文字データベースは、京都大学21世紀COE「東アジア世界の人文情報学研究教育拠点」および科学研究費補助金研究成果公開促進費188129・198092・208069・218078・228079・248061の成果物である。

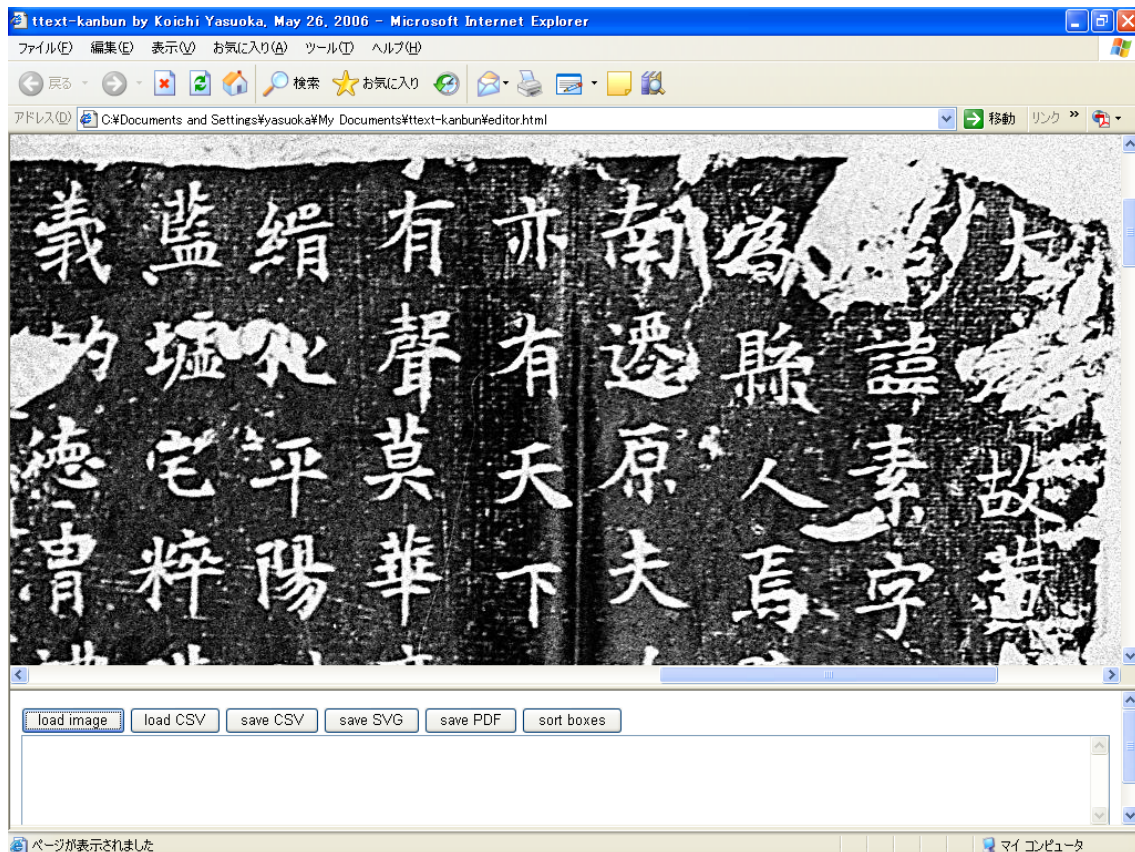


図 1 ttext-kanbun での拓本デジタル画像読み込み

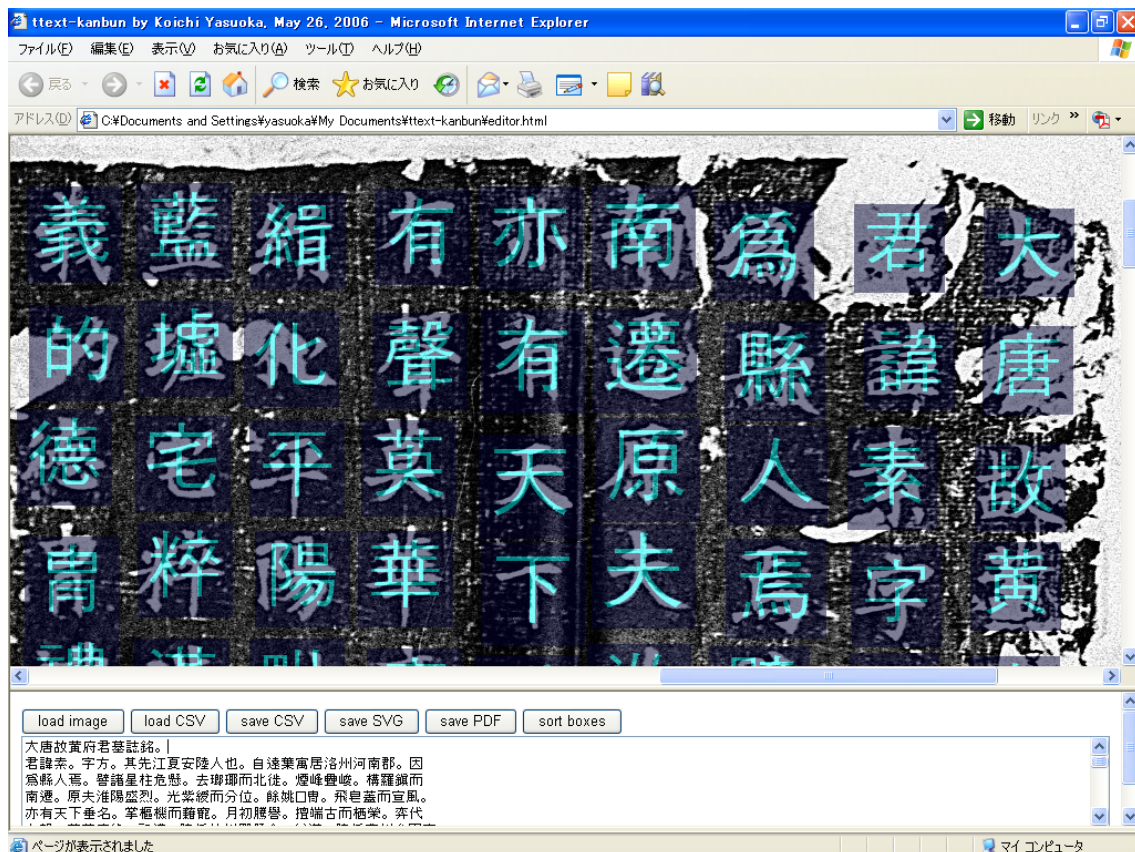


図 2 ttext-kanbun での文字切り出しおよび積文入力

2.2 文字切り出しツールの開発

拓本の撮影およびデジタル画像化に関しては、写真業者に委託することにしたが、拓本デジタル画像から文字画像を切り出して、各文字に釈文を付ける方法が問題となった。文字の自動切り出しや自動文字認識にも挑戦してみたが、残念ながら、あまり良い結果が得られなかった [1]。結局、手作業で文字画像の切り出しをおこない、それに手作業で釈文を付加するような専用ツールを開発して、拓本文字データベースのデータ入力をおこなうことにした。

筆者は当時、透明テキスト付き画像について研究しており、その作成ツールを開発していた [2]。このツール `ttext-kanbun` は、透明テキスト付き PDF や、透明テキスト付き SVG を作成するためのツールであり、Windows XP の Internet Explorer 6 上で動作する JavaScript プログラムだった。この `ttext-kanbun` に、筆者は改良を施し、拓本文字データベースのデータ入力における文字画像切り出しと、釈文の付加とを同時におこなえるようにした。この結果、実際の作業手順としては、`ttext-kanbun` に読み込んだ拓本デジタル画像 (図 1) に対して、1 字ずつマウスで囲んでいき、同時に釈文をテキストボックスに入力 (図 2) するだけで、簡単にデータ入力がおこなえる。また、釈文入力用のテキストボックスはコピー&ペーストが可能であり、事前に入手した釈文テキストを流し込んでおくことで、データ入力を、より手軽におこなえるようにしている。

2.3 WWW インターフェースの設計

拓本文字データベースの WWW インターフェース^{*2}は、2.1 節で述べた (a)(b)(c) の要件を、それぞれ直接実現する形で設計している。説明の都合上、(c)(b)(a) の順に述べる。

(c) 漢字字体データベースとしての拓本文字データベース

拓本文字データベースの基本機能は、収録されている全ての拓本に対する文字検索である。文字検索の結果は、古い時代の拓本から順に 1 字ずつ画像として表示する。例として、「京」を文字検索した場合の集字結果画面を、図 3 に示す。各文字画像上にマウスを置くと、切り出し元の拓本の標題がミニボックスに表示される。また、各文字画像をクリックすると、切り出し元の拓本の全体画像へとジャンプする。なお、拓本文字データベースの検索機能は、文字のみならず文字列も検索できるようになっている。

(b) 釈文全文データベースとしての拓本文字データベース

集字結果画面には、検索した文字あるいは文字列に対し、それを 1 文字分増減した N-gram を、生起確率順に表示している。たとえば図 3 の集字結果画面では、「京」を含む 2-gram のうち、「西京」と「京兆」が上位にあることが一目でわかるようになっている。ここで「京兆」をクリック

すると、「京兆」による集字結果画面 (図 4) へとジャンプする。あるいは図 4 の画面でさらに「京兆府」をクリックすると、今度は「京兆府」による集字結果画面へとジャンプする。

なお、孤立用例^{*3}においては、当該拓本の釈文全文がテキスト表示される (図 5)。すなわち、孤立用例に対しては、釈文の全文データベースとして、拓本文字データベースは動作することになる。

(a) 拓本画像データベースとしての拓本文字データベース

集字結果画面の各文字をクリックすると、切り出し元の拓本の全体画像へとジャンプする。図 3 の「京」の 1 つをクリックして、「唐黄墓誌銘」へとジャンプした結果を図 6 に示す。拓本の全体画像は、DjVu プラグインにより表示をおこなっており、もちろん拡大縮小も自在におこなえる。拓本の全体画像の各文字上にマウスを置くと、その文字に対する釈文がミニボックスに表示される。また、拓本の全体画像の各文字は、集字結果画面へのリンクとなっており、たとえば画像中の「京」をクリックすると、ふたたび図 3 へとジャンプする。

3. 拓本文字データベースが抱える課題

文字拓本 5,000 枚から切り出された 180 万文字の画像を、ほぼそのままの形で収録する巨大データベースであるがゆえに、拓本文字データベースは様々な問題を抱えている。この章では、拓本文字データベースが抱える問題を、2.1 節の (c)(b)(a) の順に則して解説しよう。

3.1 高頻度文字の問題

拓本文字データベース 180 万文字のうち、異なり字数は 11,800 種程度であり、1 文字を検索した際の集字結果画面には、1 ページ平均 150 字が収録されている計算になる。ただし、話はそう簡単ではない。

拓本文字データベース中で最も出現頻度の高い文字は「之」で、実に 33,300 字を超える画像が収録されている。すなわち、拓本文字データベースでうっかり「之」を検索してしまったりすると、集字結果画面には 33,300 枚もの JPEG が並ぶことになる。通常のブラウザでは、あまりの画像数にとっても耐えきれず、ハングアップすることもしばしばである。あるいは、「以」「不」「而」「子」「人」「其」「大」「有」は、いずれも 14,000~10,000 字ほど収録されており、表示の際かなりの時間を要する。

拓本文字データベースは、集字結果画面の文字画像を一定期間キャッシングしており、その意味では、ブラウザからの画像再取得が何度おこなわれても、サーバー側ではあまり問題を生じない。問題が起こっているのは、むしろブラウザ側であり、特にハングアップに対しては何らかの対

^{*2} <http://coe21.zinbun.kyoto-u.ac.jp/djvuchar> で公開中。

^{*3} 複数の同じ用例が単一の拓本のみに現れる場合を含む。

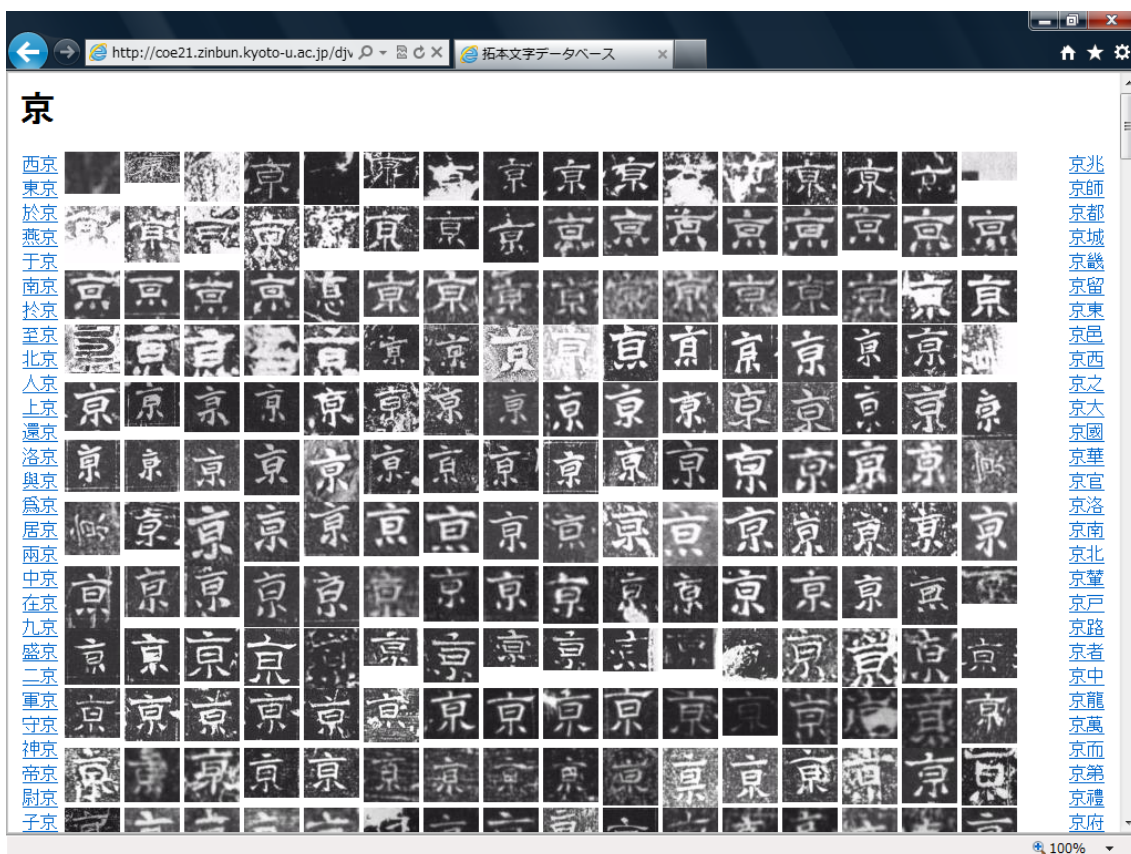


図 3 「京」の集字結果画面

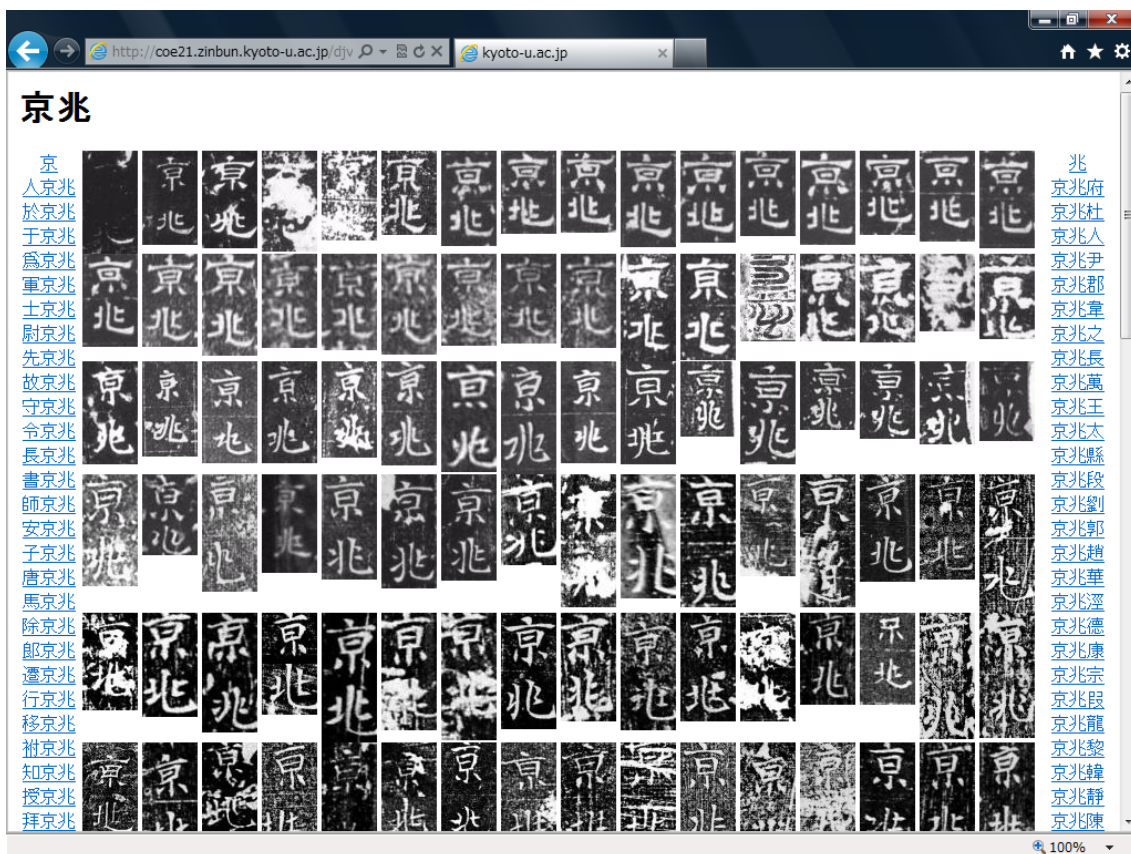


図 4 「京兆」の集字結果画面

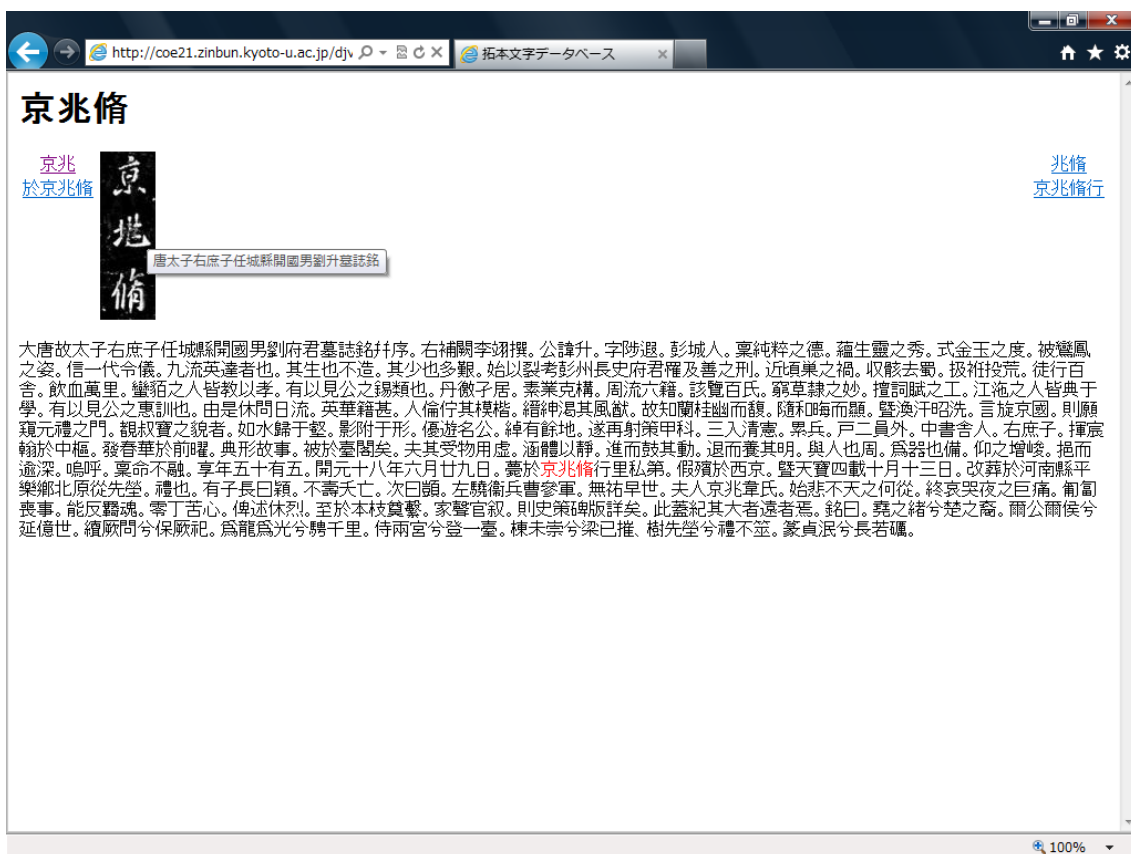


図 5 「京兆脩」の集字結果画面

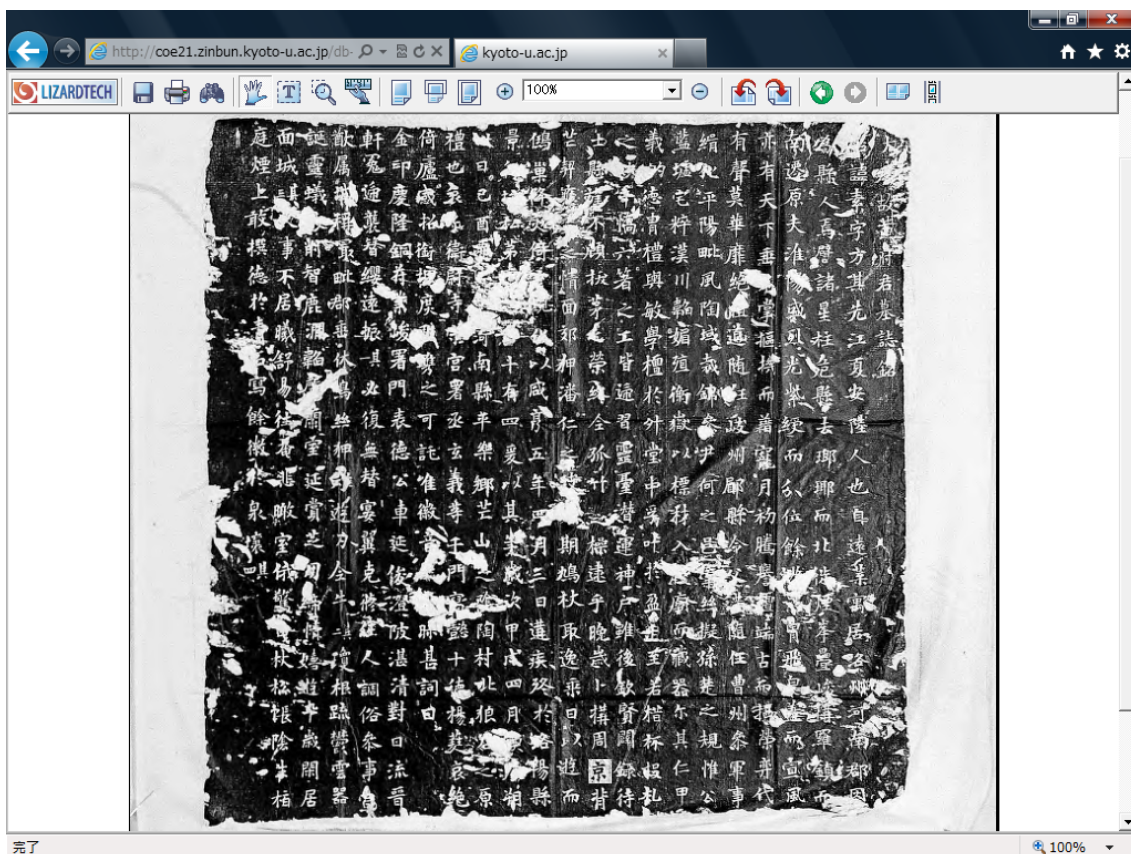


図 6 「唐黃墓誌銘」の全体画像 (画像下方の「京」が白黒反転表示)

策を講じる必要がある。ただ、たとえ 33,300 個の「之」であっても、検索結果を分割表示するのは筆者のポリシーに合わないので、画像をまとめて送るようなロジックを検討中である。

3.2 WWW クローラーの問題

拓本文字データベースでは、CHISE IDS FIND [3], [4] など他のデータベースとの連携を睨んで、全ての WWW ページに外部からリンクが張れるようになっている。集字結果画面であろうと、拓本全体画像であろうと、どこにでも外から飛び込めるようになっているのが「開かれた WWW データベース」のあるべき姿だ、という筆者の信念に基づくものである。この結果、google や Baidu など検索エンジンの WWW クローラーも、当然、拓本文字データベースの各ページに飛び込んで来てしまう。

拓本文字データベースの集字結果画面は、検索された文字あるいは文字列に対し、それを 1 文字分増減した N-gram を、生起確率順に表示・リンクしている。実は、こういうタイプのページを、検索エンジン等の WWW クローラーは、かなり苦手とするようである。何も考えずに全リンクを辿ろうとして、あちこちのページをグルグルと巡回し続けてしまうのだ。N-gram で自動生成された集字結果画面は、積文の長さを n とおいた時、 $\frac{n(n+1)}{2}$ のページ数になってしまうので、どう考えても全部まわるのは無駄なのだが、それが WWW クローラーにはわからないらしい。

やむを得ず、WWW クローラーに対しては、ある程度以上の長さの N-gram の検索をお断りしている。ただ最近では、WWW クローラーだと名乗らずに、それでも全部巡回していこうとする闇クローラーが後を絶たない。正直、困ったものである。

3.3 偽拓の問題

人文科学研究所の拓本 10,000 点には、レプリカが相当数まぎれ込んでいる。碑の一部を補刻したものから、石版刷で完全に印刷しなおしたものまで、そのレベルは様々だが、その時代の漢字字体を必ずしも伝えていないことは違いない。伊藤滋は、拓本文字データベースで公開中の漢代拓本 89 点のうち、少なくとも以下の 13 点には疑義があると指摘している [5]。

- 廩孝禹刻石
- 漢持節校尉關内侯李壇府君神道闕
- 武晉侯齊文師墓碑
- 裴岑紀功碑
- 西嶽華山廟碑
- 楊著碑
- 楊震碑
- 劉熊碑殘石
- 漢建威將軍□邈碑

- 琅邪太守朱傳頌德碑殘石
- 「■第百石」四字
- 孔融碑殘石
- 冀州刺史碑殘石

これらの偽拓 (と思われる拓本) を、どのように処遇するかについて、筆者には正直なところいいアイデアがない。積文の全文データベースとして見た場合は、それなりの価値はあるかもしれないが、拓本画像としても文字画像としても、かなりまずいシロモノである。完全に公開を止めてしまうべきか、あるいは「偽拓・時代不明」として公開を続けるか、悩ましいところである。

4. おわりに

拓本文字データベースの開発過程と現状、そして拓本文字データベースが抱えている課題について述べた。残るは、今後の展望である。

実を言えば、拓本文字データベースの「見た目」は、今後あまり変えたくない。というのも、一度「開かれた WWW データベース」を運用しはじめてしまうと、各ページの URL を変更するのは、ある意味、御法度だからだ。そのような変更をおこなうと、せっかく外部から張ってもらえたリンクを、捨ててしまうことになるからだ。

では、現在の URL とその構造を変えないという前提下で、拓本文字データベースを今後どう展開していくかだが、一つには、時間軸をもう少し可視化する、という路線が考えられる。3.1 節で示した課題を克服するためにも、古い時代の拓本から順に 1 字ずつ並べるだけでなく、何がしかの時代区分を導入する必要があるだろう。ただ、それをどういう形で実現すればよいのか。あまり「見た目」を変えずに出来るのか。そのあたりをもう少し検討してみたい。

偽拓の問題は、どこかのタイミングで全体を整理して、公開を続けるにしろ停止するにしろ、一度、何がしかの総括をおこなう必要があるだろう。ただ、偽拓の問題は、正直なところ筆者一人の手には余る。拓本の実物を専門家に見ていただくという方策を含め、実際の作業にかなりの労力が必要だ。読者諸氏も、ぜひ手助けいただきたい。

参考文献

- [1] 安岡孝一：「漢字情報学の構築」共同研究班報告，東方學報，第 83 冊 (2008)，pp.360-349.
- [2] 安岡孝一：透明テキスト付き画像作成ツールの開発，東洋学へのコンピュータ利用，第 15 回研究セミナー (2004)，pp.9-16.
- [3] 上地宏一：CHISE IDS FIND，漢字文献情報処理研究，第 6 号 (2005)，pp.163-165.
- [4] 守岡知彦：CHISE 漢字構造情報データベース，東洋学へのコンピュータ利用，第 17 回研究セミナー (2006)，pp.93-103.
- [5] 伊藤滋：碑法帖存疑 19 京大人文研石刻資料データベース考【続】，墨，第 219 号 (2012)，pp.114-117.