様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１４　日現在

機関番号：　１２１０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０３７７

研究課題名（和文）文字同定が困難な甲骨資料の画像データベース化

研究課題名（英文）Image Database of Oracle Bone Materials including Unclear Glyphs

研究代表者
鈴木　敦（suzuki, atsushi）

茨城大学・人文学部・教授

研究者番号：００２７２１０４

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：　本課題では、同定が困難な甲骨文字を含む拓本資料の画像データベースを構築した。現在、甲骨文字研究においては『甲骨文合集』が一般に利用されるが、同書の印刷品質は再版以降低下しており、文字同定の典拠とする資料として難点がある。そこで、同書の素材となった旧著録のデジタル化を行った。また、ネットワークを通じた参照利用を円滑とするため、近年人文情報学の分野で画像データベースの公開手段の標準となりつつあるIIIF方式を採用し、課題代表者が旧著録原本を所蔵しているものについては一般公開を開始した。『甲骨文合集』と旧著録の対応関係については確認ができたものから順次公開していく予定である。

研究成果の概要（英文）：　By this grant, we constructed a digital image database with IIIF protocol, for the ink-rubbings of Oracle Bone including the glyphs which are hard to identify. Today, most Oracle Bone research refers “Jiaguwen Heji” collection to check the Oracle Bone text, but its printing quality of the recent reprints are becoming worse than its original printing, and the legibility of the glyphs are becoming worse. To resolve this difficulty, we scanned the original published materials of the ink-rubbings (“Jiaguwen Heji” is a recollection of the ink-rubbings published in the past.) For the materials which copyrights are already expired, the images are already published. The database describing the relationship between the scanned images and the identifier for “Jiaguwen Heji” is now in preparation.

研究分野：　中国考古学

キーワード：　甲骨文字　データベース　著録画像データ　説文解字　小篆　ISO/IEC10646

１．研究開始当初の背景

課題開始直前に未完のまま終了したISO/IEC 10646での甲骨文字符号標準化の計画では、主に『甲骨文合集』[1]（以下、『合集』）に収録されている拓本資料に対し、人海戦術的な手法で選字作業を行い、そこから標準的な甲骨文字字形を示すフォントを作成して甲骨文字の標準文字符号としようというものであった。しかしながら、選字作業に関して網羅的な先行研究(『殷墟卜辞綜類』[2]（以下、『綜類』）あるいは『殷墟甲骨刻辞類纂』[3]（以下、『類纂』）など既に網羅的な調査がある)を無視した結果、選字作業そのものが収束しない結果に終わった。

この中で大きな問題となったのが、甲骨文字の選字基準(どのような条件を満たすものを符号化の俎上にのせるか)の不在と、それに起因する「選字作業者から出された字形に切り出しミス(本来1字でないものを1字であるかのように切り出している、など)があったり、筆画としての刻線か単なる傷かの判断が難しい場合、文字単位で切り出されているために客観的な検討ができない」という状況であった。作業者に対して『合集』のデジタル画像は提供されていた模様だが、著作権上の制限があるとのことでISOの作業グループの中でも公開されておらず、これらの作業を監査することはほとんど不可能であった。また同時に、切り出した画像について「どの拓本から切り出したか」という最低限の情報しか提示されておらず、同一拓本中に同一文字が複数見られる際に、切り出した文字を容易に確定できないことも問題であった。

この問題を解決するには、以下の二点が重要であると思われた。

- パブリックドメインではないとしても、少なくとも研究者が容易にアクセスできる甲骨拓本データベースの構築
- データベース側で文字画像を切り出して固定なデータとするのではなく、各自が切り出した位置情報を突き合わせることができるようなインタフェースの作成

２．研究の目的

甲骨文字は中華文明の文字資料としても、また、漢字の原型(古漢字)としても最古の資料である。後代の歴史書の内容を補強する程度には既に解読できているが、そのような対比ができない固有名(地名、祭名、職名など)の解釈研究は安定したとは言い難い状況である。この状況の原因の一つに、文字の同定自体が揺れる場合、資料や先行研究を調査して研究を積み重ねることができないという問題がある(『甲骨文字字釋綜覽』[4]（以下、『綜覽』）の序文を参照。)。甲骨資料には傷も多く、さらに拓本にとることによって文字を構成する刻線との区別も困難の度を増す。そのような資料および原拓に由来する問題に加え、原拓の出版部数が少なかったため、それらの収集複製資料であるところの『合集』によって研究する場合も多く、精細性が下がった資料であるために文字同定に問題が発生する状況も増えている。

甲骨文字において、文字図形が充分に鮮明であっても文字学的な同定が困難なことは、『綜覽』に整理されているが、文字学的な同定ができたとしてもISO/IEC 10646に取り込めるほど安定した代表図形を定められるかにはかなり疑問がある。

多数の甲骨資料に見える文字であれば、数個の資料で字形認識に誤りがあっても大勢に影響を与えないが[5]、少数の資料にしか見えない文字では「このような文字がある」という情報自体も絶えず検討されなければならない。この問題に対処するため、広くアクセスが可能な甲骨拓本データベースの構築が望まれる。

３．研究の方法

(1)文字同定の困難さを判断する基準の検討

文字の同定自体が揺れる場合、資料や先行研究を調査して研究を積み重ねることができないという問題に対しては、文字同定の困難さを判断する基準を検討する。即ち、

A)『類纂』が部首を決定出来なかった文字で、掲出例が 1 例のものは ISO/IEC JTC1/SC2/WG2/IRG Old Hanzi Expert Group の『説文解字』(以下、『説文』)の見出字順に排列(以下、「説文排列」)されたデータベースにどの程度採録されているか

B)『綜類』『類纂』の見出字で、どちらかしか掲出しておらず、両書の間での対応付けが不可能なものはどの程度あるか

C)『綜類』『類纂』の見出字で、掲出している出現例の数が 10 例未満のものは何項目あるか

の３項目について調査する。

(2)旧著録所収の甲骨拓本資料のデジタル画像化

併せて、「甲骨の傷や拓本の歪み等」に加え旧著録がいずれも稀覯本であることから「『合集』などに影印されたもので研究する場合も多く、精細性が下がった資料であるために文字同定に問題が発生する状況も増えている」状況への対処として、旧著録所収の甲骨拓本資料のデジタル画像化を行う。

(3)旧著録画像データベースを有効活用するためのツールの作成

(2)をデータベース化するに先立ち、データベースの具体的な使用方法を念頭に、これを有効に活用するために必要な各種ツールを検討、作成した。

(4)新たに発生した問題への対応

a) 旧著録所収の甲骨拓本資料のデジタル画像化について

本研究申請の時点では、オーバーヘッドスキャナ(線装本など資料強度に問題があり資料を平面に展開できないもの)およびフラットベッドスキャナ(洋装本など資料強度に問題がないもの)によって、研究代表者所属機関のアルバイトで実施することを想定していた。しかし、実施の過程で以下の問題が見つかった。

- オーバーヘッドスキャンを行った場合、資料を完全な平面に展開できないので、撮影したデータに対して歪み修正を行わねばならない。特に綴じ部分(いわゆるノドの部分)と、サイズが大きな資料での周辺部においては、スキャナ附属のソフトでは歪み補正が完全にできなかった。
- フラットベッドスキャナでのスキャンは資料を何度も裏返す操作があり、通常の閲覧においては問題なくても装丁が弱い資料は破損してしまう恐れがある。
- 汎用機型のフラットベッドスキャナでカラーまたはグレイスケールでスキャンした場合、データ形式を TIFF 形式としても格納されたデータは JPEG 圧縮されているため、圧縮ノイズが乗っている。600dpi 程度のスキャナでは印刷物の精細度と衝突するため、このノイズによる歪みが無視できない。モノクロスキャンであれば回避できるが、『合集』を 600dpi でスキャンすると網点とスキャン解像度が干渉して精細度が落ちることがわかっているため、これで回避することも得策ではない。オーバーヘッドスキャナあるいは USB 接続のポータブルフラットベッドスキャナはスキャン画像を無圧縮で保存することができるが、この場合のスキャン速度は著しく落ちる。

これらの問題を回避するには、資料を撮影の都度裏返さず、またある程度距離をとって撮影することで歪みを小さく抑える方法が必要である。適切な照明環境での高詳細デジタルカメラでは処理可能と考えられたが、研究代表者所属機関でそのような設備を組み

立てたとしても、アルバイトによる作業が難しいため、著作権保護期間が満了したものを業者撮影によって処理することとした。

b) データベースのインデキシングについて

本研究申請の時点では、データベースに対して『綜類』または『類纂』によるインデキシングを想定していた。これらの先行研究は、『説文』によるインデキシングでは与えられた甲骨文字が特定の『説文』見出字の小篆（以下、「説文小篆」）に対応づかないことを踏まえたものなので、特に同定に困難がある文字の扱いは『説文』では不可能と考えられたからである。しかし、初年度に中国・台湾の(文字符号専門家ではなく)甲骨研究者の意見をヒアリングした結果、文字の整理方法としては『説文』の排列順序による整理方法よりも、『綜類』・『類纂』で採用されている甲骨文字自体の字形に基づいた排列順序による整理方法評価するが、実際に文字を探そうとする場合には『綜類』・『類纂』の排列順序は専門家でも記憶できておらず、説文排列のほうが広く有用と認められる、との意見が多かった。

そこで、第二年度には

- 『綜覽』付録の「『殷墟卜辭綜類』・『甲骨文編』檢索表」により、同定困難字の『甲骨文編』[7]（以下、『文編』）における候補番号を定める。
- 対応関係を台湾の中央研究院の『文編』データベースにより検証し、対応させる拓本を定める。
- 中央研究院が公開している『甲骨文合集材料来源表』[6](以下、『来源表』)データベースにより『合集』の拓本番号に対応づける。

というプロセスで、まず『綜類』の同定困難字を『文編』に対応づけることを目指した。しかし、作業の途中で、中央研究院の古漢字関係データベースの多くが、おそらく著作権者の申し入れのために外部への一般公開が停止となり、作業が中断した(最終年度終了直前の 2017 年 3 月に一部再公開されたが、『来源表』はまだ再公開されていない)。そこで、最低限必要な『来源表』および『甲骨文合集補編』[7](以下、『補編』)付録の来源表から、内部作業用の来源表データベースの構築に取り掛かり、最終年度末までにほぼ完成することができた。『来源表』に記述されている対応関係をそのまま公開すると中央研究院と同様の懸念を抱えることになるため、旧著録および『合集』の拓本画像を実際に比較し、その検証結果という新たな知見として公開できるよう準備を進めている。

c) 古漢字の標準化動向の再燃とその対応

本課題の実施中、再度、台湾・中国による古漢字の標準化活動が再燃した。今回は説文小篆(小篆一般ではない)の標準化を先行させる計画となっていた。しかし、甲骨文字に比べれば材料が閉集合に近い説文小篆であっても、網羅的な調査を欠くことや、文字の同定基準を明確に議論しないこと、また、Variation Selector などを用いない方向性は、先の甲骨文字符号標準化計画と同一であり、同様の帰結をもたらすものと予想された。そこで、この問題への対応のため、『説文』の版本調査と字形差分析および Variation Selector を用いた字形差の取り扱いの参考実装を作成し、Unicode Technical Committee への提案を行った。

d) 人文情報学における資料画像公開の動向

計画開始当初は、同定困難な文字を含む部分画像を多数作成し、これを正当な引用の範囲として公開する予定であった。しかし、その後の甲骨文字研究の動向を調査すると、拓本にとどまらずに原骨まで遡り、文字が刻まれた領域の凹凸なども勘案して字形を論じようという研究も一部には出てきていることがわかった。これらの情報は文字部分だけ切り出すと失われてしまうので、拓本全体と文字部分のクローズアップができる仕組みが必要である。これまで、このような機能を

提供するデータベースも構築されているが、多くの場合、Web ブラウザではなく特殊な閲覧ソフトを用いるか、またはサーバ側に細かい作りこみが必要となるので、構築後の移行が難しいこと、複数の所蔵機関が同じ仕組みで提供しないために機関をまたいだ資料の比較が難しい、などの問題があった。

そこで、近年、人文情報学の分野で画像公開サービスと画像閲覧ソフトウェアを分離できる標準規格として注目されている IIIF プロトコルによる画像公開を行うこととした。

４．研究成果

(1)当初予定に沿った成果

a) 文字同定の困難さを判断する基準の検討結果

研究方法の(1)に記した作業の結果、まず A)により、説文排列に基づく文字選定では『類纂』が抽出した 151 例中わずか 2 例しか選定できておらず、『綜類』『類纂』の比較検討を行うことが妥当であることが示された。そこで B)に関する作業を進め、見出字のみで対応付け可能なものが 4 割程度という、当初推定に整合する結果を得たが、文脈などを検討しても大幅な対応関係の追加は容易でないことも分かった。そこで出現例数を直接調査すべく、C)において各項目の掲出拓本数をカウントした。自動抽出による誤差の問題は完全には解決していないが、両書とも掲出拓本数が 10 個以上の項目が 900 項未満であり、対して 1 個の項目は 1,000 項以上と見積もられた。

以上より、少数出現字の同定問題は新出資料によるものではなく、従来からある問題であり、『合集』以前の旧著録に対する網羅的な調査が必要であることが示された。

b) 旧著録所収の甲骨拓本資料のデジタル画像化

研究方法の(2)、(4)-a および-d に記した作業の結果、旧著録所収の甲骨拓本資料のデジタル画像化が、基本的に完了した。著作権保護期間を満了しており、研究代表者が原本を所持しているものについては IIIF によるインターネット公開を開始した。

(2)新たに発生した問題への対応を通じて得られた成果

a) 説文小篆の版本間字形差調査結果について

研究方法の(4)-c に記した作業の結果、説文小篆の「各種版本相互」さらには「同一版本内部」においても、どのような字形差が議論の対象になるかは研究者によって大きな揺らぎがあり、粒度が最大で 20 倍違うことが明らかとなった。このような状況で、最も細かい粒度に合わせるような符号化は危険であり、Variation Selector などを用いたある程度の階層化が必要であるという従前の主張が裏付けられた。

b) 本研究用に開発した画像処理プログラムの他の歴史的文字への適用

本研究で開発した画像処理プログラムを、ISO/IEC10646 標準化が進行中の女書資料にも適用し、現在の符号原案の問題点について指摘した。

c) 旧著録-『合集』対応データベースについて

研究方法の(4)-b に記した作業により、『来源表』および『補編』の付録のデータベース化が完了した。中央研究院のデータベースは番号のみの検索にもオンライン環境が必要であったが、本課題の中で制作したデータベースはオフラインでも動作するものになっている。

(3)成果の公表と発展の可能性

研究期間中に行った研究発表・論文発表等は、別途一覧に記す通りである。

現時点で撮影が完了している資料は全て著作権保護期間が満了しているが、原本は研究グループ外で所蔵されているものも多い。原本所蔵機関と協議の上、合意が得られたも

のから公開を開始する予定である。

旧著録-『合集』対応データベースに関しては、拓本画像によって対応関係を検証し、新たな知見として公開できるよう準備を進めていきたい。

(4)残された課題

本研究期間内に完了させる予定であった「出土例数 10 例未満の字形に関する網羅的調査」は、成果の(1)の調査の結果、1,000 個以上のサンプリングが必要となった。当初想定したよりもはるかに多い文字数が見つかったため、旧著録全体のデジタルイメージ化および拓本番号からの参照を容易にすること、文字画像領域を別情報とすることで対応を図ったが、「出土例数 10 例未満の字形を含む拓本番号の一覧」に関しては旧著録-『合集』対応データベースの公開が完了し次第処理していきたい。

[1] 『甲骨文合集』(中国社会科学院歴史研究所,1977-82)
[2] 『殷墟卜辞綜類』(島邦男,1969)
[3] 『殷墟甲骨刻辞類纂』(姚孝遂,1989)
[4] 『甲骨文字字釋綜覧』(松丸道雄，高嶋謙一,1994)
[5] 「『甲骨文編』における検索上の障害について」(鈴木敦,茨城大学五浦美術文化研究所五浦論叢,Vol.10,2003)
[6] 『甲骨文合集材料来源表』(胡厚宣,1999)
[7] 『甲骨文編』(中国社会科学院考古研究所,1965)
[8] 『甲骨文合集補編』彭邦炯,謝済,馬季凡,1999)

５．主な発表論文等

〔雑誌論文〕(計 10 件)

①鈴木俊哉、鈴木敦、菅谷克行、『説文解字篆韻譜に見える説文解字繫傳 25 巻所収文字の状況』、情処研報、CH-113No.3、1-8、2017、査読無

②鈴木俊哉、鈴木敦、菅谷克行、『＜説文解字＞小徐本の版本比較における字形差判断基準の調査』、 第 15 回情報科学技術フォーラム講演論文集、第 4 冊、15-22、2016、査読有

③鈴木敦、鈴木俊哉、『＜殷墟卜辞綜類＞における文字域排列方式の分析』茨城大学人文学部紀要人文コミュニケーション学科論集、巻 19、89-109、2015、査読無

④鈴木敦、鈴木俊哉、『＜殷墟卜辞綜類＞の部首内排列方法の分析』、情報処理学会研究報告 DD-94-6、1 巻、1-16、2014、査読無

〔学会発表〕(計 4 件)

① 鈴木敦、鈴木俊哉、『The Standardization of Ancient Chinese Characters and Attendant Problems』、第 61 回国際東方学者会議シンポジウム、2016/5/20、日本教育会館（東京都）

〔その他〕

本研究で撮影した甲骨の旧著録画像の内、「編著者の死後 50 年を経過しているもの」で「撮影原本所蔵者との、公開に関する調整が完了したもの」を公開している。今後、条件が整った物から順次追加していく。一例を挙げる。

戩壽堂所藏殷虚文字(1917)、王国維(～1927 卒)、姫佛陀(～1964 卒)

http://gyvern.ipc.hiroshima-u.ac.jp:36736/public-iipsrv/JiaGuWen/08.html

６．研究組織

(1)研究代表者

鈴木　敦　（SUZUKI Atsushi）

茨城大学・人文学部・教授

研究者番号：00272104

(2)研究分担者

菅谷克行　（SUGAYA Katsuyuki）

茨城大学・人文学部・教授

研究者番号： 30308217

鈴木俊哉　（SUZUKI Toshiya）

広島大学・情報メディア研究センター・助教

研究者番号：70311545