CHISE に基づく甲骨文字資料の電子化について

守 岡 知 彦†

CHISE の文字処理技術を用いた甲骨文字資料の電子化の試みについて述べる。

Digitization of Oracle Bones based on CHISE

MORIOKA TOMOHIKO†

This paper explains an overview of the current state of a digitization of Oracle Bones based on the CHISE character processing technology.

1. はじめに

CHISE[9] の文字処理技術を用いた甲骨文字資料の電子化の試みについて述べる。

甲骨文字は 19 世紀末に発見された中国の古代文字であり、殷周時代の古代中国を研究するための文書資料としても、この時代の中国語を研究するための言語学的資料としても、また、漢字の原義を研究するための文字学的資料としても極めて重要なものだといえる。

我々漢字利用者にとって、甲骨文字は、しばしば、最初期の漢字、漢字の祖先であり、漢字の本来の姿を表したもののように見えてしまう。この立場では、甲骨文字は漢字の一種として捉えられ、(現代) 漢字との対応関係が強調され、広義の異体字処理の枠組の中で処理されるべきものとなる。しかしながら、甲骨文字には (現代) 漢字との対応がつかない文字も多数ある。また、現代なら複数文字で書くような、複数音節や複合語を 1 文字で書く「合文」という現象や、上下左右反転可能な場合があるなど、『漢字原理』の枠組に収まらない性質を持っている。しかしながら、部品の組合せによる造字システムや異体字の存在、現代字と極めて似通った簡略化や部品の組合せも見受けられ、後の漢字に受け継がれたさまざまな特徴を備えているのも確かである。

一方、筆を使って甲骨文字を紙に書いた作品を作り続けている一部の書家たちの多くにとって、おそらく、甲骨文字は前者の立場で捉えられるべきものなのであろう。この場合、甲骨文字は、事実上、古代文字というよりも、漢字の書体の一種というように捉えられているといえる。

このような甲骨文字の多面性を損なうことなくうまく記述するためにはどうすれば良いかという問題は、単に、古代文字の符号化という問題に留まるのではなく、書かれた時代における視点の復元と現代における視点の齟齬をどのように考えるかということであり、現代も読まれているような古典作品・文献にも共通するような問題だといえる。

† 京都大学人文科学研究所
Institute for Research in Humanities, Kyoto University

2. 目　　標

甲骨文字に関する資料としては、一次資料としての甲骨、および、その写真 (図 1)・拓本 (図 2) 等があり、また、これらのメタデータとしての甲骨の索引や文字単位に整理した索引 (図 3) がある。また、甲骨に書かれた文章の釈文や研究論文の類がある。

多目的利用が可能なオントロジーを作るという観点からいえば、比較的学説に依存しにくい部分に立脚して骨組みを作り、その上にさまざまな解釈のレイヤーを載せて行くのが望ましいと考えられるが、こうした観点に立って、まず、一次資料、および、そのメタデータの電子化から行うことにした。また、第一目標として、甲骨文字オントロジーを実現するという観点から、文字として十分に検索・利用可能な必要最小限のメタデータを付けることを目指した。

前述のように、甲骨文字には漢字に対応する部分としない部分が存在する。このため、対応する漢字が存在する場合には異体字処理の延長線上で対処可能であるが、対応する漢字が存在しない場合には漢字を介して甲骨文字を検字することはできない。また、対応する漢字が文字列になる場合 (合字) があり得る。伝統的には、説文解字における部首 (説文部首) に基づいた分類・排列が行われてきたが、分類不能な文字も多数存在してしまう。このように、漢字との対応関係では表現し得ない甲骨文字をどのように扱うかということは甲骨文字オントロジーを作成する上で重要な課題

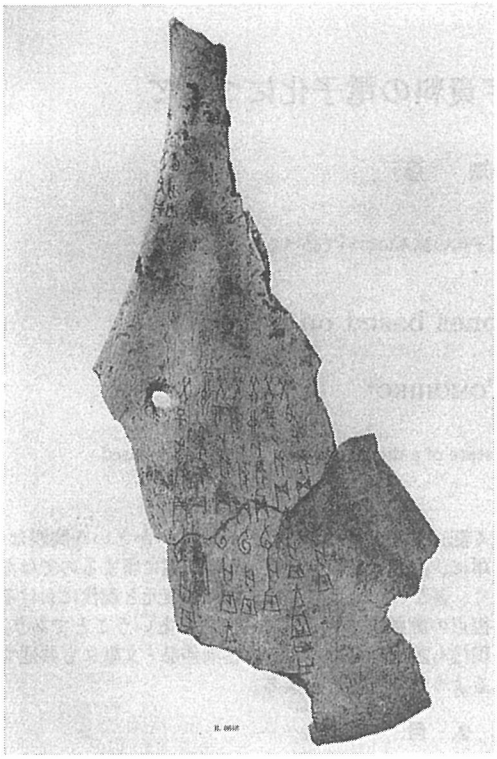


図 1　卜骨（甲骨文字が書かれた獣骨）の写真

である。

また、甲骨文字の項目からその記述に関する一次資料である甲骨の写真や拓本がひけることも重要である。

3. 入力作業

3.1 画像データ化

入力作業の効率化を考えれば、まとまった紙媒体の資料があることが望ましい。そのため、我々はまず「京都大学人文科学研究所所蔵甲骨文字」の図版册（拓本と写真☆）と文字索引篇（図 3）をアルバイトを使いスキャンし画像データ化した。

3.2 甲骨画像の整理

「京都大学人文科学研究所所蔵甲骨文字」の図版册にはほぼ全ての甲骨の拓本と一部の甲骨の写真が収録されている。そして、各甲骨には編号（以下、「甲骨編号」と呼ぶ）が振られている。甲骨編号は

甲骨編号 := 接頭辞 甲骨番号 接尾辞
接頭辞 := S. | B.
甲骨番号 := dddd（但し、d は 0〜9 の数字）
接尾辞 := a | b

☆ 拓本は全ての甲骨に対して存在するが、写真は一部に対してのみ存在する

図 2　卜骨の拓本

という風になっている。ここで、接頭辞は材質を表しており、卜甲（亀の甲羅）の場合には S.，卜骨（獣骨）の場合には B. となる。また、甲骨番号は接頭辞に依らないユニークな番号となっている。

この甲骨番号（甲骨編号）は研究者の間で京都大学人文科学研究所所蔵の甲骨を参照するために使われており、この番号を指定することで該当する甲骨の拓本・写真等の画像を表示できることが望ましい。しかしながら、図版册の 1 頁に複数の甲骨が記載されており、このままの状態では甲骨番号で画像を検索することができない。そのため、図版册の各頁をスキャンした画像データを各甲骨毎に切り出す作業を行った（当初、この切り出し作業の機械化を試みたが、完全に自動化することができなかったため、結局、アルバイトに頼った）。

各甲骨は甲骨編号で管理することとし、切り出した各甲骨の画像データに対して、

拓本の場合　rubbings/甲骨編号.tif

写真の場合　photos/甲骨編号.tif

という風な、甲骨編号を元にしたファイル名を付けて保存した。

3.3　文字索引の入力

甲骨文字を CHISE の文字処理技術の枠組で扱うためには、各甲骨文字に対する文字定義が必要となる。また、エディタ等で表示するためにはフォント（グリフ情報）も必要となる。我々はこれらの情報を「京都大学人文科学研究所所蔵甲骨文字」の文字索引篇（図3）の情報を元に作成することにした。

3.3.1　文字索引篇の構造

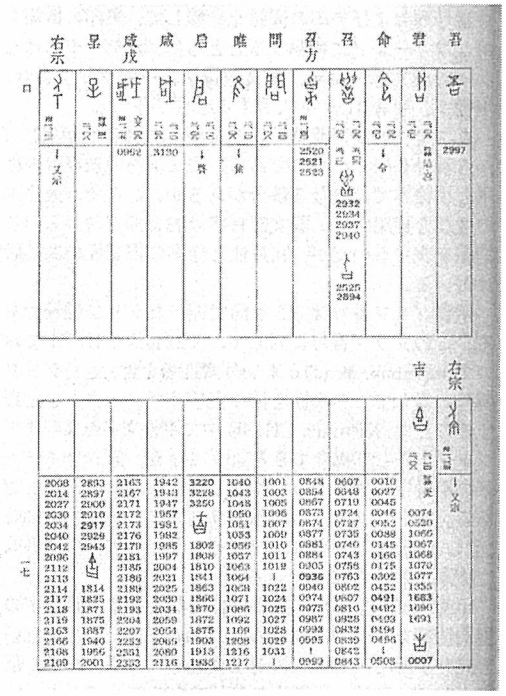


図 3　文字索引篇の例

文字索引篇の各頁（図3）は上下2段で構成されており、各段は罫線で区切られた12行からなる。各行は、罫線の欄外上部（第0フィールド）、罫線内の第1フィールド、第2フィールドからなる。甲骨文字は第1フィールド上部に書かれる。対応する現代字（隷定字、釈字）が存在する場合、第0フィールドに書かれる。第1フィールド下部の右側の漢数字は「甲骨文編」、左側の漢数字は「続甲骨文編」の編数と葉数を表す。第2フィールドのアラビア数字は 3.2 節で述べた甲骨番号を表す。また、顕著な異体字が存在する場合、第2フィールドに記載される。また、第2フィールドの末尾には「⇒」もしくは「→」という矢印が記載されることがある。「⇒」は、「京都大学人文科学研究所所蔵甲骨文字」本文篇において新しく読みかえたことを示し、「→」は

(1)　第1フィールドに甲骨文字、および、「甲骨文編」、「続甲骨文編」の編数と葉数を記入したものは、重文☆および参看の場合

(2)　甲骨文字が空欄で、「甲骨文編」、「続甲骨文編」の編数と葉数のみを記入したものは、両著の誤釈と判断されるもの

(3)　全て空欄のものは、「京都大学人文科学研究所所蔵甲骨文字」本文篇での釈字に異同があるものを示す。また、第0フィールド、第1フィールドがともに空欄の行は継続行である。

上段の、奇数頁の場合は左側、偶数頁の場合は右側の欄外にこの頁に現れる説文部首が記載されている。また、下段には頁番号が漢数字で記載されている。

3.3.2　文字データの入力

文字索引篇に記載された情報の内、通常の文字データとして入力可能なものをアルバイトを用いて入力することにした。この作業には、Microsoft Excel を用い、第1列に頁番号、第2列に現代字、第3列に「甲骨文編」の編葉数、第4列に「続甲骨文編」の編葉数、第5列に甲骨番号列および矢印の情報（第2フィールドの情報）を入力するようにした。

現代字の項目には、隷定のために無理矢理作ったような字や、合文の一部において隷定されていない甲骨文字が現れたりする。また、日常使わない漢字も多数現れるので、作業者が入力できなかった字は「？」を入れることとした。

第5列の甲骨番号列は、甲骨番号をスペースで区切ったものを入れることにした。また、内容に関して理解していない作業者でも入力できるように、また、校正作業のことも配慮し、なるべく、文字索引篇の情報をそのまま入力することとした。よって、甲骨番号列と矢印の情報は構造化されておらず、原文に見えたように入力されることになる。また、原文において改行があれば、入力データも改行することにした。但し、異体字が現れた場合は、異体字が現れた場所で改行することにした。

3.3.3　甲骨文字の切り出し

文字索引篇の第1フィールド上部、あるいは、第2フィールド中に現れる甲骨文字字形を文字毎に切り出した（当初、この切り出し作業の機械化を試みたが、認識がうまく行かない場合があり、完全に自動化することができなかったため、結局、アルバイトに頼った）。切り出した甲骨字形画像は、甲骨文字が第1フィールド上部の代表字形の場合、

頁番号-行番号. 拡張子

第2フィールドに現れる異体字形の場合、

☆ 異体字

頁番号-行番号_列番号. 拡張子

というファイル名を用いた。ここで、頁番号は 10 進 4 桁の番号（4 桁未満の場合は先頭に 0 を詰めた形式）である。また、行番号は 10 進 2 桁の番号（1 桁の場合は先頭に 0 を詰める）で、頁内の行番号を 00 ～ 23 のように上段・下段を通算したものである。また、列番号は、10 進数の番号で、各行毎に、第 2 フィールドに甲骨異体字形が現れた順に 1 から数え上げた番号である。このようなファイル名を用いることにより、見落しや誤りの際の影響が少なくなり、また、校正もしやすくなった。

3.3.4 フォント作成

前節のように切り出した約 4000 字の甲骨文字字形を加工して、128 ドットの BDF 形式のビットマップフォントを作成した。また、これを縮小して、48 ドットのフォントも作成した。フォント利用環境の制約を考えれば、1 つのフォントに含まれるグリフ数は 65536 個以下であるのが望ましいといえ、グリフの番号もまた 0 ～ 65535 の範囲に収まる方が良いといえる。前節のファイル名のままではスパースな大きな番号になってしまうため、☆ 1 から始まる番号を機械的に振り、これをグリフ番号とした。

3.4 文字索引の加工・校正

3.3.2 節で入力した Excel データを TAB 区切り形式の Unicode テキストとして書き出し、これを元に XEmacs CHISE[10] で加工・校正作業を行っている（図 4）。

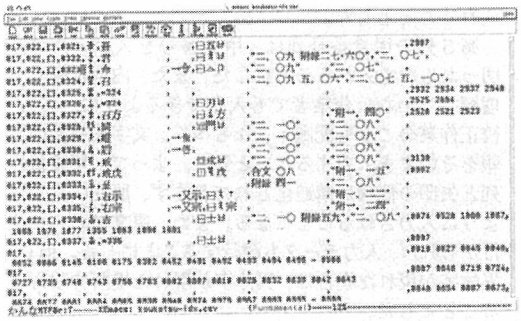


図 4 加工・校正された甲骨索引データ

3.3.2 節で入力したデータは、第 1 列に頁番号、第 2 列に現代字、第 3 列に「甲骨文編」の編葉数、第 4 列に「続甲骨文編」の編葉数、第 5 列に甲骨番号列および矢印の情報が入っているが、加工・校正済みデータでは、TAB 区切りではなくコンマ区切り (CSV) を用いるとともに、

（1） 頁番号（10 進 3 桁の自然数）
（2） 説文部首番号（10 進 3 桁の自然数）
（3） 説文部首字
（4） 甲骨グリフ番号（10 進 4 桁の自然数）
（5） 甲骨字
（6） 現代字
（7） リンク（矢印の）情報
（8） 甲骨 IDS
（9） 「甲骨文編」の編葉数
（10） 「続甲骨文編」の編葉数
（11） 甲骨番号列

というフィールドからなる形式を用いることにした。よって、元データでは第 5 列に混在して入っていた甲骨番号列および矢印の情報を分離して、矢印の情報を第 7 フィールドに移動するとともに、元データにはなかった説文部首番号、説文部首字、甲骨グリフ番号、甲骨字、甲骨 IDS を入力する必要が生じる。

ここで、説文部首番号は説文解字の部首を順番に 1 から番号を振ったものである。説文部首の並びは大徐本と小徐本では異なる部分があるが、ここでは大徐本の並びを採用する。説文部首字は説文部首番号と同じ情報を表すものだが、冗長性と作業の視認性の為に設けている。

甲骨グリフ番号は 3.3.4 節で述べたグリフ番号である。このグリフ番号に対して、XEmacs CHISE において `=zinbun-oracle` という素性名を付け、各グリフ番号に対してこの素性を持つ文字オブジェクトを定義することで、XEmacs CHISE 中で甲骨文字を文字オブジェクトとして操作できるようになる。第 5 フィールドの甲骨字はそうやって作成した文字オブジェクトである。なお、ファイルに保存する際には `&ZOB-dddd;`（但し、dddd は甲骨グリフ番号）という実体参照形式で表現することにしている。

甲骨 IDS（図 5）は複数の部品からなる甲骨文字の構造情報を IDS (Ideographic Description Sequence) 形式[4] に準じた形式で記述したものである。但し、部品としては前述の甲骨字の使用を認めたものとなっている。

この加工・校正作業にはアルバイトも用いたが、この場合は、文字索引篇の頁内に存在する情報のみを入力・校正することとし、存在しない情報に関しては空欄とした。即ち、甲骨グリフ番号と甲骨字が入力対象であり、説文部首番号、説文部首字、および甲骨 IDS のフィールドはとりあえず空欄とし、改めて、著者がこれらの項目を埋めるという作業手順をとることとした。

現時点では、アルバイトによる作業が第 2206 字まで進んでおり、この内、第 1856 字まで説文部首番号と説文部首字の入力が完了している。

☆ 頁番号を 8 ビット、行番号を 5 ビット、列番号を 3 ビットで表現すれば、16 ビットで表現することは可能であるが…（今思い付いてしまった（^_^;）。

4. CHISE でのデータ表現

3.4 節で加工・校正した CSV 形式の甲骨索引データは各フィールドに対して文字素性[9] を対応付けることによって、CHISE の文字定義に変換することができる。こうした方法によって、CHISE の文字オントロジーに甲骨文字の情報を統合して管理している。

4.1 頁 番 号

甲骨索引データの第 1 フィールドに記載された頁番号の情報を表現するために文字素性 *zinbun-oracle-page* を設けている。

4.2 説 文 部 首

説文部首の情報は説文部首を表す文字素性 *shuowen-radical* で表現する。この素性の値は（大徐本における）説文部首番号（自然数）である（将来、大徐本と小徐本などの版の差異の情報を記載する必要があればドメイン識別子[9] を用いることとする）。

4.3 現代字との関係

甲骨索引データの第 6 フィールドに記載された甲骨文字に対応する現代字の情報は文字素性 *<-Oracle-Bones* で表現する。この素性は関係素性[9] の一種であり、この素性を持つ甲骨文字に対応する現代字のリストを値にとる。

この素性は逆関係素性 *->Oracle-Bones* と関係素性対を構成しており、XEmacs CHISE で甲骨文字に *<-Oracle-Bones* 素性を設定すると、自動的にその値に設定された各現代字に対して逆素性 *->Oracle-Bones* が設定される。これにより、現代字から対応する甲骨文字をひくことができるようになる。

隷定や釈字には複数の学説や出典が存在し得るが、この場合にはドメイン識別子を用いることにする。

4.3.1 合 文

ところで、甲骨文字には現代では複数文字で表現するようなものを 1 文字で表した「合文」というものが存在するが、この場合、現代字（隷定字、釈字）は 1 文字で表現できず、文字列となってしまう。

こういう場合、対応する現代字を文字列オブジェクトにするというのもひとつの方法であるが、この場合、甲骨文字における文字概念を記述するという観点では問題があるように思われる。

そこで、文字処理の枠組に閉じた形でも合文を扱えるようにするために、値として文字のリストを取り、その文字の連なりで文字列を表現する *ideographic-combination* 素性を設けることにした。この素性を用いれば、文字列を便宜上の文字オブジェクトとして表現することができる。例えば、対応する漢字が「中子」であれば、

```
((ideographic-combination ?中 ?子))
```

という便宜上の文字オブジェクトとなる。

合文はこの仕組みを用いて、1 つの文字オブジェクトとして表され、*ideographic-combination* 素性の値によって、文字列としての情報も保持でき、結果的に、文字としての性質と文字列としての性質の双方を表現することができる。また、文字素性を追加することで他のメタデータも表現可能である。

4.4 甲 骨 IDS

甲骨文字も漢字と同様に、複数の部品が組合わさった文字（以下、「複合（甲骨）文字」と呼ぶことにする）が少なくない。こうした複合甲骨文字に対して部品の組合せ構造（漢字構造情報）を記述しておけば、対応する現代字がなくても（あるいは、説文部首が判らなくても）、部品を指定することによってその部品を含む甲骨文字を検索することができる。このように、甲骨文字においても漢字構造情報は有用であるといえる。

甲骨文字の漢字構造情報を記述する形式は漢字の場合と同様に幾つかの形式が考えられるが、我々は現代字（漢字）の場合と同様に、IDS (Ideographic Description Sequence)[4] 形式に準じた形式を用いることにした。これは IDS と同様に IDC (Ideographic Description Characters)[4] をオペレータとする前置形式であるが、部品として甲骨文字を使用することを認めたものである。これを「甲骨 IDS」（図 5）と呼ぶ。

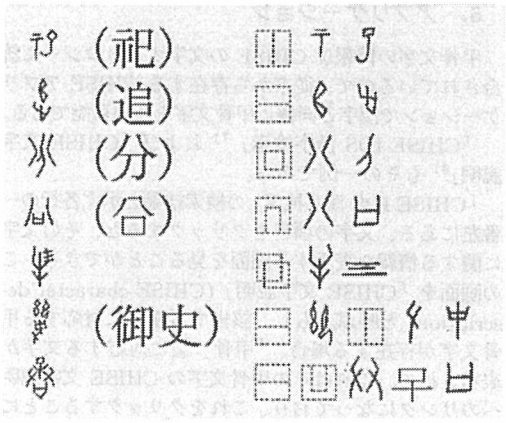


図 5 甲骨 IDS の例

3.4 節で述べたように、甲骨索引データの第 8 フィールドには、この甲骨 IDS が記載できるようになっているが、CHISE では、漢字の場合と同様に、これを構文解析して、S 式で表現した情報を *ideographic-structure* 素性[8] の値として記載することにしている。これにより、「CHISE IDS 漢字検索」[7] などのシステム・サービスがそのまま利用可能となった。

4.5 甲骨番号列

甲骨索引データの第 11 フィールドに記載された甲骨番号列は、甲骨文字の出典情報と看做すことができ、出典情報を表す文字素性 *sources* に対応付けることに

した。

sources 素性は出典に対するユニークな識別子のリストを値として持つことになっている。このため、単に甲骨番号のリストを設定するのではなくて、各甲骨番号に zob1968= という接頭辞を付け、zob1968=1234 のような出典識別子を用いることにした。このような手法を用いることにより、他のカタログ等に対しても、同様に、固有の接頭辞を与えることにより、さまざまな出典の情報を混在して収録することが可能となる。

4.6 異体字

異体字に関しては、とりあえず、見出し字を便宜的に抽象文字オブジェクトのように扱い、顕著な異体字として出されたものをその子オブジェクトと看做すことにし、

見出し字（親）<-*denotational* 異体字（子）

という文字間の関係として記述した。しかしながら、見出し字もまた例示字形の一種であることを鑑みれば、本来は本当の抽象的な文字概念を親とし、見出し字はその子オブジェクトとして記述すべきである。よって、これは今後、他のカタログや UCS での符号化の進展を考慮し、これらとの関係が十分に記述できるように配慮しつつ、再整理されるべきものと考えている。

5. アプリケーション

甲骨文字の情報は CHISE の文字オントロジーに統合されているので、従来から存在する CHISE アプリケーションで漢字と同様に甲骨文字を利用可能である。「CHISE IDS 漢字検索」[7] および「CHISE 文字説明」[9] もその一例である。

「CHISE IDS 漢字検索」の検索結果を示す各行の一番左にある、文字の項目をクリックすると、その文字に関する情報を表示する画面を見ることができる。この画面を「CHISE 文字説明」(CHISE character description) と呼ぶ。もし、該当する漢字に対応する甲骨文字が存在する場合、「甲骨」欄に対応する文字が表示される。これはその甲骨文字の CHISE 文字説明へのリンクになっており、これをクリックすることによって甲骨文字の情報を見ることができる（図 6）。

甲骨文字の CHISE 文字説明の「sources」欄は出典である甲骨の拓本画像へのリンクになっており、これをクリックすることで拓本画像を見ることができる。

また、漢字の場合と同様に、「漢字構造（解字）」欄は部品へのリンクとなっている。また、「この部品を含む漢字を探す」ボタンをクリックすると、この甲骨文字を部品として含む甲骨文字を検索することができる（図 7）。

6. 関連研究

甲骨文字の符号化に関しては、台湾中央研究員や文字鏡などでの例があるが、ここでは、最近、

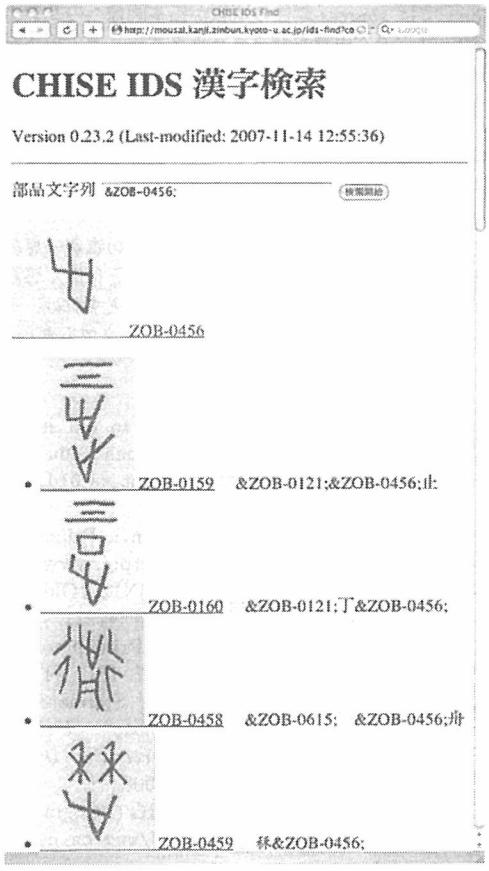
図 6 甲骨文字の CHISE 文字説明の例

ISO/IEC JTC1/SC2/WG2/IRG (Ideographic Rapporteur Group; 以下、IRG とする)[5] で進行中の UCS[4] に甲骨文字を収録するための作業に関して簡単に述べる。

6.1 IRG における標準化作業

IRG は UCS における漢字の符号化を担当しているグループであるが、最近は現代漢字とは別に、先秦期の古代漢字の符号化について議論されており、中でも甲骨文字に関する議論や作業は他の古代漢字よりも進んでおり、収録字の選定ルール、包摂・分離に関する原則、甲骨文字のデータベースの形式といったことに関して草案を伴った具体的な議論が行われており、実際に、甲骨文字のデータベース化が行われているようである。著者は IRG での甲骨文字の符号化作業に参加しておらず、IRG の WWW サイト[5] に上がっている文書を見ただけであるので、その実情に詳しくないことをあらかじめお断りしておく。また、本件に関しては、小形克宏氏の記事[6] に詳しくまとめられているので、興味のある方はこちらを一読されることをお勧めする。

IRG による甲骨文字のデータベースの形式に関しては “Old Hanzi Principles and References” という文書で規定されている。この文書は現在のところ Version 3[3] が最新のようであるが、Version 2[2] は IRG のサイトで読むことができる。



これによれば、データベースは

(1) ID
(2) 代表字形
(3) オリジナルの拓本字形
(4) 出典
(5) 時期・時代
(6) 地域・場所
(7) 材質
(8) 説文部首
(9) 説文部首番号
(10) 隷定字
(11) 対応する現代字（UCS の符号位置）
(12) 包摂可能な形状
(13) 注

という欄からなり、11, 12, 13 番目の欄はオプションである。同一の形状の文字であっても出典が異なれば別のレコードに分けるようになっており、文字毎にまとめられていない。これは、おそらくは、現時点では何をもって同じ文字とするかを検討している段階であるということと、単純な表形式の構造で表現したいということからなされたことであろうと思われる（これは妥当な判断であると思われる）。

本研究の甲骨索引データと比べた場合、前述のように出典毎に別レコードになっていることと、隷定字と対応する現代字が分離されていることが異なるといえる。時期・時代、地域・場所、材質の欄の有無も表面的には異なるが、本研究の甲骨索引データの場合、出典となる出土した甲骨の属性となり、本質的には異ならないといえる。但し、本研究では現在の所、出土した甲骨に関するメタデータの電子化が進んでいないため、時期・時代、および、地域・場所の情報が欠けている。材質に関しては、甲骨編号の接頭辞で表現されているので、卜甲、卜骨といった種類が判る。ちなみに、IRG のデータベースの場合、時期・時代欄は「商」、素材欄は「甲骨」という粒度の荒い情報しか入っておらず、実質的には、現状でもあまり差異がないといえるかも知れない。

UCS における甲骨文字の符号化が最終的にどのようなものとなるかは現状では判らないが、おそらくは、なんらかの包摂規準によって包摂され（そして、幾つかの顕著な異体字は別の符号化文字として分離され）、そうして確定したレパートリを説文部首順に並べた符号位置を振られたものとなるのではないかと思われる。

ここで、説文部首に基づく整理を行うことに対して批判が存在する。[6)1)11)] 説文解字が著された後漢代と甲骨文字が使われた時代では文字概念に差異があり、甲骨文字においては同じ字であったものが説文解字では別字になっている例や、説文部首で分類できない甲骨文字が存在するからである。甲骨文字は現代の漢字から見れば、漢字の祖先であり、最初期の漢字という風に看做すことができるが、実際には現代の漢字とは直接対応しない甲骨文字も多数あり、対応する文字であっても、現代とは文字概念が異なっている場合も多々あり得る。また、造字法などの点で漢字と共通する文字システムを持っているといえるが、合文のように現代の漢字原理とは異なる性質も有している。☆ よって、甲骨文字自体の性質に根ざした分類、即ち、甲骨部首のようなものを考えるべきであるという立場は一理あるといえる。しかしながら、コンピュータ上のデータベースを前提とするならば、実の所、どういう並びであるかはあまり重要なことではないといえる。説文部首と甲骨文字の類も必要があれば両方の情報を付加すれば良いのであり、両立可能であるといえる。また、文字を探すという観点でいえば、部首法に基づく検字よりも、本論文で提案する甲骨 IDS のような、漢字構造情報を利用した手法の方が便利であるといえ

☆ そもそも、漢字原理は説文部首による帰納的な説明を逆に演繹的にフィードバックした結果生じたような面もあるように思われる。

る。また、CHISE 的な手法を採るのであれば、異なる包摂規準・文字概念に基づく表現も両立可能であるといえる。

符号化という観点でいえば、誰のための（何のための）符号化であるかということが重要な問題であるといえる。UCS は合理的な符号化のために符号化対象となる文字に関する学術的知見を必要とするといえるが、その結果できた符号は必ずしもその専門領域のためだけのものではないといえる。即ち、現実の情報交換のために符号化されるべきだといえるのである。そういう観点で考えた場合、甲骨文字のような古代文字はそもそも工業標準に馴染まないという考え方もあり得るが、今日のようにさまざまな領域で文書の電子化が進んでいることを考えれば、甲骨文字符号の標準化には一定の意義はあると思われる。ここで、甲骨文字符号の想定される利用者を考えてみれば、甲骨文字自体の研究者の他にも、甲骨文字で書かれたテキストを利用する歴史学者や言語学者などの研究者、甲骨文字に関わる論文のメタデータ、あるいは、研究者以外によるライトユースなどが存在するのではないかと思われる。おそらく、人口や利用件数でいえば、後に挙げたもの程多くなって行くと思われる。情報交換用符号としては、おそらく、メタデータによる利用が十分に行えることが重要となるのではないかと思われるが、いずれにせよ、甲骨文字自体の性質（殷代や周代における文字観念）も重要な問題ではあるが、現代の情報交換用符号とては現代における甲骨文字の利用実態の観点も必要となるのではないかと思われる。

また、データベースという観点でいえば、甲骨文字からその出典となる出土した甲骨に辿れること、即ち、一次資料の電子化ということが重要であり、次に、それが容易に検索でき、さまざまな利用ができることが望ましいといえる。学説に依存する部分に関しては複数の学説の情報が両立できることが望ましい。こうした観点でいえば、IRG のデータベースは出土した甲骨の情報を持っているという点で最低限の要件を満たしているということはいえるが、甲骨の画像情報そのものは入力されていないので、文字が置かれた文脈情報を得ることができないという点では問題があるといえる。

7. おわりに

CHISE の文字処理の枠組に基づいて甲骨文字オントロジーを実現する試みについて述べた。現在、約 1900 文字分の甲骨文字の情報が CHISE の文字オントロジーに統合されており☆、これらは XEmacs CHISE や「CHISE IDS 漢字検索」などで利用可能となっている。後者にリンクされた「CHISE character description」では *sources* 素性の情報が拓本画像へのリンクになっており、甲骨番号をクリックすることで拓本写真を見ることができる。

今後は、*ideographic-structure* 素性の入力を進めるとともに、他のカタログや将来の UCS での甲骨文字符号を統合し、より多角的で実用性の高い甲骨文字オントロジーに発展させたいと考えている。

☆ 但し、*ideographic-structure* 素性が付与されているのはその一部である

謝　辞

本研究を行う上で、富山大学人文学部の森賀一惠教授から貴重なご意見・ご助言を頂いたことに深く感謝する。なお、本論文における誤りがあるとすれば、それらはすべて著者の責であることはいうまでもない。

参考文献

1) Atsushi Dr. SUZUKI. Input to old hanzi expert group. `http://www.cse.cuhk.edu.hk/~irg/irg/irg29/IRGN1346-Input_to_Old_Hanzi.pdf`, Sep 2007. IRG N1346.
2) Old Hanzi Group. Old Hanzi Principles and References (Version 2). http://www.cse.cuhk.edu.hk/~irg/irg/irg27/IRGN1271OldHanziPrinciplesV2.pdf, Dec 2006. IRG N1271.
3) OldHanzi Group. Old Hanzi Principles and References (Version 3), Jun 2007. IRG N1336.
4) International Organization for Standardization (ISO). *Information technology — Universal Multiple-Octet Coded Character Set (UCS)*, March 2003. ISO/IEC 10646:2003.
5) ISO/IEC JTC1/SC2/WG2/IRG (Ideographic Rapporteur Group). `http://www.cs.cuhk.edu.hk/~irg/`.
6) 小形克宏. UCS における甲骨文字収録の意義と問題点. 全国文献・情報センター人文社会科学学術セミナーシリーズ、京都大学学術情報メディアセンター 第 79 回研究セミナー, pp. 151–173, Mar 2007.
7) 守岡知彦. CHISE IDS 漢字検索. `http://mousai.kanji.zinbun.kyoto-u.ac.jp/ids-find`.
8) 守岡知彦. CHISE 漢字構造情報データベース. 全国文献・情報センター人文社会科学学術セミナーシリーズ、京都大学学術情報メディアセンター 第 78 回研究セミナー, pp. 93–103, Mar 2006.
9) 守岡知彦. 文字オントロジーに基づく文字処理について. 情処研報, Vol. 2006, No. 112, pp. 25–32, Oct 2006. 2006-CH-72.
10) 守岡知彦, 師茂樹. 文字素性に基づく文字処理. 情処研報, Vol. 2004, No.58, pp. 53–60, May 2004. 2004-CH-62.
11) 鈴木敦. Old hanzi 審議状況への意見. http://www.cse.cuhk.edu.hk / ~irg / irg / irg29 / IRGN1346-Attachment-B.pdf, Sep 2007. IRG N1346 Attachment.