

文字情報サービス環境 CHISE

連載 | 第1回

—素性の集合に基づく文字表現—

京都大学人文科学研究所 もり おか とも ひこ 守岡 知彦

はじめに

文字は古来よりさまざまな社会制度や文化の基盤であり、公的な文書、定型化された書類、会計記録、歴史、経典、教科書、実験記録、学術書、小説、詩、脚本、書簡や日記など、さまざまなテキストが書かれてきた。本や書類以外にも、ポスターや商品の包装、テレビ画面上のテロップなど、見渡せば日常生活のあちこちに文字が書かれている。

こうした文字の利用はインターネットやモバイルデバイスの普及などによってさらに加速したといえる。たくさんの人々が日々SNSやメッセージングアプリ、メール等で会話するようになり、ネット上のコンテンツを読むようになった。但し、ネット上を流れる文字はそれ以前の紙に書かれた文字と異なる性質がある。それは文字を符号化する必要があるということだ。

文字を符号で表現できるのは、文字に与えられた番号(符号位置)が同じか違うかだけである。言い替えれば、文字符号を作る前に符号化したい文字のどの性質を表現したい(区別したい)のか、そしてどの差異を捨象して良いか、そういったことをその文字を利用する人達にとってなるべく違和感がなく、社会制度や言語政策に齟齬がないように定義する必要がある。

文字は一般に書体や書き手などによってさまざまな書かれ方をされる。ちなみに、紙に書かれたり印刷されたりディスプレイに表示された文字の具体的な形状のことを字形(glyph image)と呼ぶ。一般にある抽象文字(符号化された文字)はさまざまな形状で書かれた無数の字形に対応している。例えば、A, A, Aのように。

ラテン文字の場合、どの書体で書かれてもこれらがAの大文字であることが判るし、正書法の存在する書記言語の場合、必要な文字の種類もあらかじめ判っている。つまり解釈共同体が共有する文字概念のセットがあらかじめ判っているのも、それらの文字概念を抽象文字として符号化すれば良いだけである。

しかしながら漢字の場合、文字数が多い上に、「體」と「体」や「畧」と「略」のように同じ音価・字義の組合せ(即ち、古典

中国語の形態素)が複数の全く異なる形状で書かれたり(異体字)、「藝」の略字と「芸亭」の1字目のように複数の異なる形態素(音義)が同じ字体で書かれたり(別字衝突)、「大」と「犬」は別の字だが「類」「類」の場合は左下の部品をどちらで書いても音義が同じ(同字)、しかし、「伏」と「伏」は別字になるなど、仮に漢字の抽象的な形(字体)を符号化するとしてどの差異を区別しどの差異を捨象すれば良いか簡単には判らない。漢字は時代や地域によって形状や音価や字義(形音義)が変化してきたため、異体字関係は基本的に時代・地域依存であり、異なる地域・時代の漢字を字体でまとめた漢字のレパートリーを作れば必然的に少なくない数の別字衝突が生じてしまうからだ。

また、漢字は複数の部品を組み合わせで別の文字を作り出すことができるため、原理的には無限に存在し得るといえるが、一つのテキストで必要となる字種は数千~1万字前後であり、6万字程度あれば四書五経をはじめとする中国の(儒教系)古典も概ね表現可能であるので、青天井のコード領域を用意することは経済的ではない。しかし、仏教教典中の文字、地名や人名、方言文字、出土文字を考慮すれば実はこれでは足りない。特に、出土文字に関しては考古学的調査や盗掘された文物の発見等によって未知の文字が見つかることがあり、まだ研究中の文字の場合、音義が判らない場合もある^{※1}。

そういう意味では、漢字には未解読の文字としての側面があるといえる。つまり、解釈共同体の判断が期待できないケースがあり、言い替えれば、文字概念=抽象文字を適切に定義しづらいケースが存在するということだ。

今回のテーマとなるCHISE(Character Information Service Environment)^{※2}とは、このような問題を解決するために著者が中心となって1999年頃から開発を続けている文字情報処理のためのシステムである。CHISEでは文字(特に漢字)に関するさまざまな知識を『文字オントロジー』という機械可読な

※1 大漢和辞典やUnicode等にも多数の音義未詳字(あるいは、一方が判らないもの)が収録されている。

※2 <https://www.chise.org>

データの集合（データベース）として記述することによって文字符号に依存しない文字処理を実現している。ここでは、その考え方や概要について述べていく。

素性の集合に基づく文字表現

CHISEでは『Chaonモデル』と呼ぶ方法によって文字を表現するようになっている。これは汎用符号化文字集合に依存することなく自由に文字を表現するために我々が提案しているもので、表現したい文字に関する知識（文字の性質の集合）の機械可読な描写によって文字を表現し操作する方法である。

これを理解するための例として、電話（音声だけ）で漢字を伝えることを考えてみる。例えば、私の名前を伝えるために「わたしのなまえは／モリオカ／です」と言った場合、音声情報／モリオカ／だけでは「守岡」という情報を伝えることができない。多くの場合、私の名前は「盛岡」や「森岡」と表記されてしまう。そこで、「『守る』の／モリ／に『岡山』の／オカ／です」とか「うかんむりに寸の／モリ／に…」などという風に漢字を説明する訳である。このように文字に関する性質や用例を用いて文字を指し示そうというのがChaonモデルの基本的な考え方である。

Chaonモデルでは、文字を説明するための要素（文字の性質や用例など）を『文字素性』（Character Feature）と呼ぶ。文字素性としては、部首、画数、部品の組合せ方に関する情報（漢字構造情報）、発音、意味、用例、その他文字処理で必要となる各種情報などが考えられる。このようにさまざまな種類の文字素性が考えられるが、CHISEではこれらを

- CCS素性** 文字符号の符号位置を示すための素性
- 関係素性** 他の文字との関係を示すための素性
- 構造素性** 他の文字を部品とした時の部品の組合せによる構造を示すための素性
- 一般素性** 上記以外の素性

に分類して扱う。

Chaonモデルに基づく文字素性を実際にコンピュータの上に載せるためには、幾つかの方法が考えられるが、一番簡単なのは1つの文字オブジェクトを『連想配列』と呼ばれるデータ構造で表現することである。連想配列はキーと値の対の集合からなり、Lisp（S式）の『連想リスト』、JSONの『オブジェクト』は連想配列の一種である。また、これは素性名と素性値からなる表としても表現可能である。

例えば、「字」という字に関する情報を右の表1

表1

素性名	素性値
康熙字典部首	39（子部）
部首内画数	3画
総画数	6画
漢字構造	「宀」の下に「子」
UCS	U+5B57
JISX0208	0x3B7A（27区90点）
大漢和辞典	06942
大字典	02267

のように表現することができる^{※3}。ただし、このままでは十分な機械可読とはいえないので、素性名や素性値の機械可読形式を定める必要がある^{※4}。

例えば、CHISEのS式表現の場合、康熙字典部首は1～214の部首番号で表現され、素性名としてideographic-radicalを用いている。また、部首内画数はideographic-strokes、総画数はtotal-strokesという素性名で表現され、それぞれ素性値には画数の値が格納される。漢字構造はISO/IEC10646（＝Unicode）で定義されているIDS（Ideographic Description Sequence）形式に基づき、その構文木である入れ子状のリストを素性値に格納し、素性名としてはideographic-structureを用いている。よって、これらの4素性の組は

```
((ideographic-radical.39) ; 子
 (ideographic-strokes.3)
 (total-strokes.6)
 (ideographic-structure ?[宀?子]))
```

という連想リストで表現することができる。

一方、インターネット上でこの種の情報を交換するためにRDF（Resource Description Framework；資源記述の枠組）が標準化され、メタデータやLinked Data、セマンティックWeb等の基盤として用いられている。

RDFはデータモデルであり、具体的なデータ形式としてはXMLを用いたRDF/XMLやJSONを用いたJSON-LD、RDFグラフを簡潔にプレーンテキストで表現するために作られたTurtle（Terse RDF Triple Language；RDF/Turtle）などが存在する。

CHISEは、内部ではRDFと同様な表現力を持つ独自の有向グラフストレージのConcordを用いているが、RDFもサポートしており、RDF/Turtle形式での出力機能やJSON-LDベースのWebAPI、RDF用の問い合わせ言語であるSPARQLを利用した検索サービスも提供している。（続く）

※3 CHISEにおける実際のデータは <https://www.chise.org/est/view/character/字> で見ることができる。また、<https://api.chise.org/character/v1/get-info?character=字> でJSON-LD形式のデータが得られる。

※4 cf. 「CHISEのデータ形式」<http://git.chise.org/~tomo/character/chise-format.pdf>

文字情報サービス環境 CHISE

連載 | 第2回

—漢字構造記述—

京都大学人文科学研究所 もり おか とも ひこ 守岡 知彦

1 はじめに

多くの漢字は複数の要素を組み合わせて字を構成している。例えば、「林」という字は「木」を左右に並べた形をしているし、「雲」という字は「雨」の下に「云」を並べた形をしている。

ここでは、漢字の構成要素のことを「部品」とし、複数の部品の組合せからなる漢字の構造（部品の組合せ方）のことを「漢字構造」と呼ぶ。そしてその情報の記述を「漢字構造記述」と呼ぶことにする。

漢字構造記述の標準形式としてはISO/IEC 10646（及びUnicode；本稿ではこれらを総称してUCSとする）で定義されたIdeographic Description Sequence (IDS) 形式が普及している。

2 IDS

IDSはLisp言語におけるS式と同様な前置記法的一种で、前置演算子であるIDC (Ideographic Description Characters) (図1) の後に2個ないしは3個の部品 (Description Components (DC)) をとる。例えば、2個の部品を左右に並べる前置演算子「」を使って「村」の漢字構造を「木寸」と書くことができる。

IDSでは部品として、

- UCSで符号化された漢字
- UCSで符号化された部首・部品用文字
- U+FF1F (全角の「?」)
- 外字
- IDS

を用いることができ、部品としてIDSを用いることによって、入れ子状の記述が可能である。例えば、「溪」の漢字構造を「溪」ハ夫」と書くことができる。

IDSは漢字構造記述を文字列として扱ったり、文字列としてパターンマッチングしたりするには便利であるが、部品としてIDSが現れる場合の処理はそのままでは少し面倒である。よっ

て、より込み入った処理を考えれば、IDSを構文解析した結果の構文木として扱う方が便利であるといえる。

そこで、CHISE文字オントロジーでは内部表現としてはLisp言語におけるS式を用いている。IDSはLisp言語におけるS式と同様な前置記法なので、その構文解析は極めて容易であり、オペレータの前とオペレータがとる最後の部品の後に括弧を付ければS式となる。

CHISE文字オントロジーではChaonモデルに基づき文字素性の集合として文字を表現しているが、漢字構造記述はこの文字素性の1つと看做することができ、前回の記事で述べたように、素性名としてはideographic-structureを用いている。

このCHISEのideographic-structure素性の値に格納される部品もまた文字素性の集合で表現される文字オブジェクト（これを「部品オブジェクト」と呼ぶことにする）であるが、これはUCS

Ideographic description characters	
<i>These are visibly displayed graphic characters, not invisible composition controls.</i>	
0 	2FF0  IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO RIGHT
1 	2FF1  IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO BELOW
2 	2FF2  IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO MIDDLE AND RIGHT
3 	2FF3  IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO MIDDLE AND BELOW
4 	2FF4  IDEOGRAPHIC DESCRIPTION CHARACTER FULL SURROUND
5 	2FF5  IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM ABOVE
6 	2FF6  IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM BELOW
7 	2FF7  IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LEFT
8 	2FF8  IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER LEFT
9 	2FF9  IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER RIGHT
A 	2FFA  IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER LEFT
B 	2FFB  IDEOGRAPHIC DESCRIPTION CHARACTER OVERLAID

図1 Ideographic Description Characters

で符号化されていない文字も利用可能である。良く使われている部品オブジェクトの集合としては台湾中央研究院で作られたCDP外字などがある（なお、現在ではこのCDP外字の多くがUCSに収録されている）。また、グリフ作成やインターネット上で共有の容易さからGlyphWiki^{*1}も良く使われている。

3 多粒度漢字構造モデル

一方、部品として複数の異なる包摂粒度を持つものを用いれば、複数の部品の組合せで構成される漢字の各部品の包摂範囲を示すことで、その漢字の包摂範囲を示すことができるといえる。これを『多粒度漢字構造モデル』と呼ぶ（図2）。CHISE文字オントロジーにおける漢字構造記述は基本的にこのモデルに基づいている。

UCSの統合漢字は複数の字体（グリフ）を包摂した抽象文字ということにはなっているものの、過去に各国・地域で用いられていた文字符号との互換性を考慮して包摂除外を行ったり、誤って重複符号化したケースもあり、また、部首・部品用文字と統合漢字との関係も明確に規定されていないため、IDS中の部品として利用可能なUCSの文字の包摂範囲には曖昧性がある。これに対し、CHISEでは漢字や漢字部品に対して、それぞれ、抽象文字や字体などの複数の包摂粒度の文字オブジェクトを定義することで、包摂粒度情報が付いた漢字構造記述を実現している。

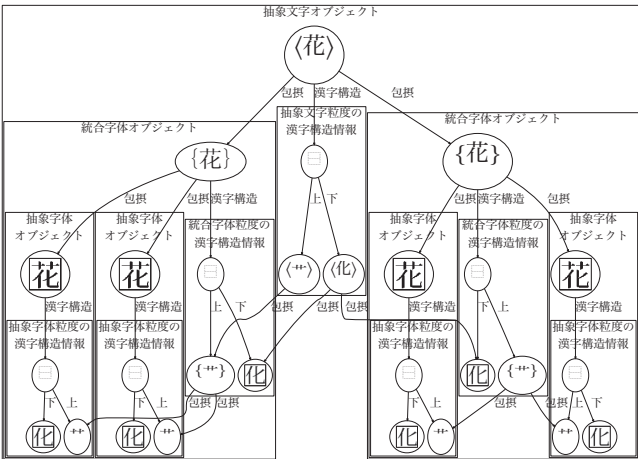


図2 多粒度漢字構造モデルの概念図（花）

※ 1 <https://glyphwiki.org/>
明朝体の漢字字形を登録・管理し、皆で自由に共有することを目的としたWebサービス。UCSに漢字を追加提案する際にも良く使われる。

※ 2 CHISE-IDSをCHISE文字オントロジーへ読み込む際、既にideographic-structure素性値があるものに関してはCHISE-IDS側の記述を無視するようにしている。

4 CHISE漢字構造情報データベース

「CHISE漢字構造情報データベース」（CHISE-IDS）はUCS統合漢字やいくつかの外字集合の漢字に対してIDS形式で漢字構造記述を行ったものである。2022年8月末の時点でUCS収録済みの漢字のIDSを92750件、全部でのべ212197件のIDSを収録している。

CHISE-IDSはIDS形式のデータベースであるため、実体参照形式で記述された外字などを除けば、大部分の部品はUCS統合漢字であり、包摂粒度の情報を持たない。しかしながら、これらは一度CHISE文字オントロジーを用いて、各部品をUCSの抽象文字として解釈し対応するUCSの抽象文字オブジェクトのideographic-structure素性値に張り付けるとともに、Unicode3.2の例示字形が存在するものに関しては、それらに対応する字体粒度のオブジェクトとしても解釈し対応する字体粒度オブジェクトのideographic-structure素性値に張り付けることで、1つの粒度なしIDSから複数の粒度付きIDSを生成するようにしている。また、粒度なしIDSではうまく曖昧性を解消できないケースに関しては、CHISE文字オントロジー中で直接記述することで対処している^{*2}。

5 漢字の部品

「林」を構成する部品「木」は「林」という字が「木」に関係する文字であることを示唆する一方、「林」と「木」の字音が全く異なることから「林」の部品「木」は「林」の音価に関する情報を担っておらず、もっぱらその意味カテゴリーを示す部品（これを「意符」もしくは「義符」という）であるといえる。

同様に、「雲」を構成する部品「雨」も意符であるが、もう一つの部品「云」は「雲」の音価に関する情報を担っている（これを「音符」もしくは「声符」という）。大部分の漢字は意符と音符の組合せで構成されたもの（これを形声字という）である。漢字は意味を表す文字（表意文字）というイメージが強いが、実の所、表音文字的な性格も強く、表音文字としては音節文字の一種と考えることができ、また、この音価が古典中国語の形態素を示すことから「表語文字」といわれる。

漢字の部品には「意符」や「音符」の他、純粹に見掛け上の形を示すもの（これを「形符」という）も存在する。「意符」や「音符」には元となった漢字があり（文字単体としては廃れてしまったり、歴史上用例はない場合もあるが、そうした場合でもその部品を持つ漢字の集合の共通的な性質として、部品に対応する文字概

念が共有されていて、その字義や字音が忘れられず意識されている)、その部品を持つ漢字を知らなくてもその意味カテゴリーや大まかな字音を想像することができる。

だが、「春」の部品「𠂔」の場合、同じ部品を持つ「奉」「奏」「秦」「秦」「春」などに見比べても、字義的な共通性も字音的な共通性も字音的にも思い浮かばない。しかし、「𠂔」という部品を持つ漢字が複数存在し一定の生産力を持っていることが判る。ここで、「意符」や「音符」のように元となった漢字の音義等の情報の一部を保持した部品のことを「機能的部品」と呼び、「形符」のような音義等の機能的なつながりが存在しないような見掛け上の部品のことを「皮相部品」と呼ぶことにする。

6 複数の漢字構造

「旗」の漢字構造を考えた場合、その分解の仕方によって、

- 方曰一其
- 方其

という2つの記述を考えることができる。この時、前者の漢字構造に現れる部品「方」「其」からは「旗」の音義との関連性を見出すことができないのに対し、後者の漢字構造の場合、「𠂔」は単体字「𠂔」の変形した形で旗に関連する意符、「其」は音符と分析することができ、形声字であることが判る。

このように、部品が入れ子状に組合わさった漢字の場合、その分解の仕方が一つに定まらない場合がある。この時、前者のような見掛け上の漢字構造を「皮相漢字構造」、後者のような機能的部品の組合せで構成される漢字構造を「機能的漢字構造」と呼ぶことにする。大抵の場合、皮相漢字構造は、変換規則を用いることによって、機能的漢字構造から自動生成することができ、CHISEでは文字オントロジーのインストール時にこの変換処理を行っている。また、自動変換が難しいケースに関しては、CHISE-IDSのオプションフィールドに皮相漢字構造を記載したり、CHISE文字オントロジー中で直接記述することで対処している。

7 「CHISE IDS 漢字検索」

CHISE文字オントロジー中の漢字構造記述を用いて漢字を検索するためのサービスとして2005年4月から「CHISE IDS漢字検索^{※3}」というWWWサービスを提供している。

これは、検索窓(「部品文字列」)に探したい漢字に含まれる部品を入力して検索を実行すると、指定した部品を持つ漢字の



図3 CHISE IDS漢字検索の検索結果の例

リストが表示されるというものである。指定する部品は1つでも良いし複数でも良い。複数の部品を指定した場合、指定した部品を全て持つ漢字が表示される(図3)。

検索結果は1行1文字で、ある検索結果を部品として持つ漢字が存在する場合、それらがインデントして表示される。ちなみに、ある検索結果を部品として持つ漢字を部品としても持つ漢字が存在する場合、以下、再帰的にインデントして表示される。なお、検索結果が非常に多い場合、このインデント表示や再帰的な検索は省略される。

インデント表示される場合、検索結果の各行の左側をクリックすると、その文字の詳細情報が表示される。インデント表示が省略された場合、検索結果に現れた各文字をクリックするとその文字の詳細情報が表示される。

(続く)

※ 3 <https://www.chise.org/ids-find>

文字情報サービス環境 CHISE

連載 | 第3回 (最終回)

—漢字の字体用例と古字書—

京都大学人文科学研究所 もり おか とも ひこ
守岡 知彦

1. はじめに

文字は一般に書体や書き手などによってさまざまな書かれ方をされる。そして、漢字の場合、文字数が多い上に、微小なデザイン差に加え、字体差(異体字)も少なくない上、別字衝突(同形(類形)異字)も存在する。漢字は時代や地域によって形状や音価や字義(形音義)が変化してきたため、異体字関係は基本的に時代・地域依存であるといえる。

この結果、どの差異を区別して捨象すれば良いか簡単には判らなくなった一方、文字は言語的コミュニケーションのためのツールでもあるので、少なくともある共同体の中で誰かが書いた文字がちゃんと読める必要があるといえる。

解釈共同体においては文字の同一性(言い替えれば、どの差異を区別しどの差異を捨象すれば良いかの判断)の規準が共有されているはずである。また、おそらく標準的な字体も存在しているはずである。こうした各時代・地域における漢字の解釈共同体の文字意識(の変遷・差異)を明らかにするためには、それぞれの典型的な文字資料の字体を収集し、異なる形の文字が実際に書き分けられているのかどうかを調査することが重要である。

一方、各時代において良く使われる文字や注意すべき文字を収集・整理し、音義や異体字関係などを解説した漢字字書が作られており、特に前近代に作られたものを古字書と呼ぶ。漢字字書の記述は同時代における現実の漢字使用実態を表すものとはいえない(また、玉篇が説文解字を引用しているように、前時代のものを反映しがちである)が、漢字に関するある種の規範意識を知る上で重要な資料である。また、康熙字典や大漢和辞典(における古字書の引用)はUnicode収録漢字の重要な典拠のひとつ^{※1}となったことで、現代の文字符号にも影響を与えている^{※2}。

本稿では、歴史的漢字字体用例のデータセットとして「漢字字体規範史データセット」(Hanzi Normative Glyphs dataset; HNG dataset)、そして古字書のデータセットとして「平安時代漢字字書総合データベース」(HDIC) について紹介する。

2. HNG

漢字字体規範史データセット(HNG)^{※3}は、時代や地域毎の漢字字体の標準の存在とその変遷を明らかにすることを目的に構築された漢字のグリフデータベースである。その前身は石塚晴通氏が30年程前から作成を続けてきた字体資料(「石塚漢字字体資料」と呼ぶ)である。

「石塚漢字字体資料」は紙カード(図1)で整理されていたが、15年程前から電子化が開始され、現在ではその紙カードを撮影した画像とそこから切り出された代表字形の画像と関連するメタデータが<https://gitlab.hng-data.org/HNG/>上のGitリポジトリとして公開されている。また、「漢字字体規範史データセット単字検索」(略称「HNG 単字検索」)^{※4}という検索サービスも用意されている(図2)他、CHISE文字オントロジーとも統合されており、CHISE IDS漢字検索やCHISE-wiki等のCHISE関連Webサービスからも閲覧可能である。

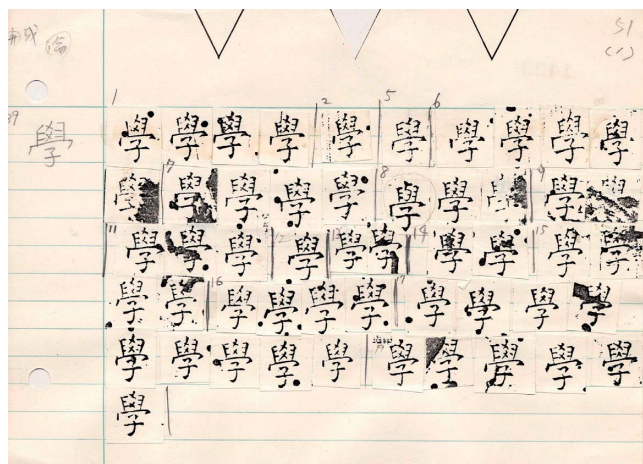


図1 石塚漢字字体資料の紙カード(開成石經論語: 學)

※1 整理やソーティングなどにも用いられた。

※2 戸籍統一文字や文字情報基盤等の日本の行政用文字符号も影響を受けている。

※3 <https://www.hng-data.org/>

※4 <https://search.hng-data.org/>

これらは互いにリンクされており、CHISE-wikiの文字表示画面の一番上にある「HNG-data」ボタンを押せば対応するHNG単字検索の画面に遷移し、HNG単字検索の一番上の「見出し字」、「大字典番号」、「大漢和番号」の値部のリンクや、各用例画像、用例数をリンクするとそれぞれ対応するCHISE-wikiの頁に遷移する。このように、両者はリンクを辿ることで往き来でき、必要に応じて好きな検索手段・viewerを選択可能である。

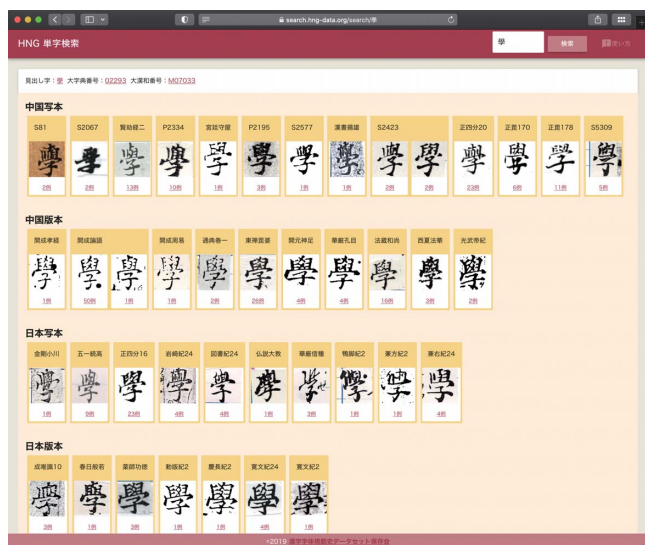


図2 HNG単字検索

3. HDIC

HDIC^{※5}は、漢字字体史研究と漢字字書編纂史研究に資することを目的に、池田証壽氏のグループが開発している中国・日本の漢字辞書（古辞書・古字書）のデータベースである。

HDICは中国の古辞書である原本玉篇（梁・顧野王撰、543年成、現存2,087字）、宋本玉篇（宋・陳彭年等撰、1013年成、約22,800字）、龍龕手鏡（遼・行均撰、997年成、約26,000字、高麗版）、日本の古辞書である篆隸万象名義（空海撰、9世紀初、約16,000字）、新撰字鏡（天治本、昌住撰、10世紀初、約21,000字）、類聚名義抄（図書寮本、11世紀初、約3,600項目／観智院本、12世紀後半、約32,000項目）を対象に、掲出字画像データベースと全文テキストデータベースを構築することを目指しており、現在は宋本玉篇と篆隸万象名義、新撰字鏡、観智院本類聚名義抄のテキストデータベースがGitHub上で公開されている^{※6}。

また、検索サービスとしてHDIC Viewer^{※7}が公開されている（図3）。



図3 HDIC Viewer（天治本新撰字鏡「字」）

4. CHISEを介した統合

HNGとHDICはそれぞれ独立したデータセットであり、独自の検索サービスを持っているが、CHISEとも連携・統合されており、例えば、CHISEの文字情報表示用の頁（CHISE-wiki）上においてHNGやHDICの情報がCHISEの情報と並べて表示される（図4）。

HDICには漢字字体史研究のための歴史的漢字字形データセットとしての側面があり、HNGが欠いている字書の掲出字の情報を補完するものとして、両者の連携は重要な課題のひとつであったが、CHISEを介することでこれを実現することができた訳である。

また、HDICは古字書の注文（辞書記述）のテキストデータを持っているが、の中には異体字関係に関する記述も存在する。CHISEでは現在試験的に宋本玉篇の注文を自動解析してCHISE文字オントロジーの関係素性や既存の関係素性の典拠データを追加することを行っており、将来的には他の古字書にも広げることを計画しているが、これによりHDICの注文内容の機械可読化がなされるだけでなく、CHISEのデータの典拠を容易に調べることもつながるといえる。

HDICが収録する玉篇や篆隸万象名義などは説文解字を引用しているが、CHISEは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の東方学デジタル図書館で公開されて

※5 <https://hdic.jp/>

※6 <https://github.com/shikeda/HDIC>

※7 <https://viewer.hdic.jp/>

← 字種：祭

→ [HDIC] 篆隸萬象名義：祭 子裔反。祭也，至也，薦也，享也，祀也。〔祭入：古文。〕

→ [HDIC] 宋本玉篇：祭 (反切)子滯切。薦也、祭祀也。又側界切。周大夫邑名。

→ [HDIC] 新撰字鏡：祭 子裏反。祀也，至也，薦也。祭 子裏反。薦也。鄧字也。在邑部。

→ [HNG] 中国写本：

→ [HNG] 中国版本：

→ [HNG] 日本写本：

→ [HNG] 日本版本：

→ 甲骨文字：

→ Small-Seal：

→ 説文小篆：

→ [HDIC] 篆隸萬象名義 篆書掲出字：

→ formed：𩇑 祭 祭 祭

図4 CHISE-wiki での表示例
(<https://www.chise.org/est/view/character/祭より>)

いる3種類の説文解字の掲出字(説文小篆)データを収録している^{※8}。この説文小篆データには全文画像へのリンクが付与さ

れており、このリンクを辿ることによりすぐに元の版本画像を見ることができる。よって、HDIC(中の説文引用箇所)からCHISEを介して説文の全文画像に飛ぶことができ、3種類の説文の記述とHDIC中の説文からの引用文を比較することもできる。

5. おわりに

これまで全3回の記事を通じて、CHISEのアーキテクチャと漢字構造記述、そして、今回の記事では漢字情報の基礎となる歴史的漢字字体用例のデータセットである「漢字字体規範史データセット」(HNG)と古字書のデータセットである「平安時代漢字字書総合データベース」(HDIC)について紹介した。

本稿ではCHISEの概要や機能、収録するデータの極一部しか紹介することができなかったが、漢字に関する知識の機械可読な記述について少しでもご関心を持って頂けたなら幸いである。
(了)

※8 この見出しの座標データは広島大学の鈴木俊哉氏に提供して頂いたものである。

新刊 令和3年度税制改正対応

効率とコンプライアンスを高める
e-文書法 電子化早わかり

参考資料満載！

- 電子帳簿保存法 取扱通達解説(趣旨説明)
- 電子帳簿保存法 一問一答
- お問合せの多いご質問
- 電子帳簿保存法 法的要件認証制度 (JIIMA認証)

公益社団法人
日本文書情報マネジメント協会
法務委員会 編

令和4年2月15日 発行
B5判 298ページ
ISBN 978-4-88961-019-2
価格 3,300円(税込)

◆ お問合せ・お買い求め

公益社団法人 日本文書情報マネジメント協会 (JIIMA)

<https://www.jiima.or.jp/> 「JIIMAの活動」→ 出版物・販売物 より