

GlyphWikiによる観智院本『類聚名義抄』字形の再現

劉 冠偉 (北海道大学・日本学術振興会)

李 媛・池田証壽 (北海道大学)

本研究は国語学国文学の研究における重要文献の観智院本『類聚名義抄』を対象に、漢字字形共有データベース GlyphWiki によるその写本字形を再現することを目標とする。筆写体である字形を可能な限り字体上の特徴を残し、書写の影響を捨象して活字化を行う。本稿は GlyphWiki で再現した字形とその利用についての詳細を報告する。

Recreations of Glyphs in Kanchi'inbon Ruiju Myōgi shō with GlyphWiki

Guanwei Liu (Hokkaido University / Japan Society for the Promotion of Science)

Yuan Li / Shoju Ikeda (Hokkaido University)

This study focuses on glyph creation of undefined Chinese characters in *Kanchi'inbon Ruijumyogisho* 観智院本類聚名義抄 and aims to build a glyph list comprising the full headwords in the dictionary using GlyphWiki. GlyphWiki is community-driven and stores structural information on Chinese characters. It creates and displays undefined Chinese characters in character-coded sets used in digitising Chinese text resources. It can generate font files and images from the glyph stored on the sever. A test version of our glyph list is available on the GlyphWiki.

1. はじめに

国語学研究資料のデジタル化が進んでおり、学術データベースの作成・公開が多く見られる。筆者らの所属する研究室で推進している平安時代漢字字書総合データベース (HDIC) ¹ プロジェクトは、近時刊行の高精細カラー版の観智院本『類聚名義抄』によって本文解読の精度を高めた上で、全文テキストデータベースを近日中に公開する予定である。古辞書が収録した掲出字には、次の三つの特徴がある。

(1)古辞書は膨大な数の漢字を収録し、異体字も多数。古辞書にしか記録がなく、実用例が存しない字も多々ある。

(2)古辞書のほとんどが古写本として伝存されているため、筆や書きぶりの影響で筆画の細かい部分を認識、抽象化することが困難である。

(3)先行文献の引用や転写の過程で誤写の箇所が多い^[1]。

研究や教育の目的による古辞書における漢字字形の利用は、前述の三つの特徴に直面し、古辞書の原本画像を参照しなければならない状況がほとんどである。原本に書かれた字形を画像データとして扱うことが格段に容易になった。しかし、古辞書の原本画像は、所蔵機関の権利関係等により、二次利用の面において制約が課せられている。

加えて、古辞書の古写本における漢字字形は、難読の内容が多い。これは、同形異字・類形異字の存在、時間・空間の隔たりによる標準字体の変遷等の影響が関係する^[1]。

研究資料としてのオープン性を高めると同時に解読した利便性の高い古写本字形表現をする方法が求められる。筆写体である字形を可能な限り字体上の特徴を残して、筆写の影響を捨象して活字化を行う。そのデータをインターネットで共有する。

そこで、漢字字形共有データベース GlyphWiki による観智院本類聚名義抄字形の再現することを考えついた。具体的に、次の三つのステップで進行する。

① GlyphWiki を用いて、HDIC の公開用インターフェース HDIC Viewer に字形作成プラットフォームを追加する。(4.2)

② 観智院本『類聚名義抄』において認められる外字を作成する。(5.1)

③ 同システム GlyphWiki で Unicode に既存する掲出字に対して、所在情報を付与することで、すべての掲出字字形の表現することを試みる。(5.2)

なお、観智院本『類聚名義抄』に関する作字作業の概要を李・劉・池田 (2020) で指摘したが、本稿はその後の進行状況及び改善等を報告するものである。

¹ 北海道大学文学研究院言語科学講座池田証壽研究室
<https://hdic.jp/>

2. 対象資料

2. 1 観智院本『類聚名義抄』について

『類聚名義抄』は国語学国文学の研究において基本参考文献として各種の論著や注釈書に例証として引用され、『日本国語大辞典』（小学館）のような大型国語辞書にもその用例が数多く取り上げられている。『類聚名義抄』の諸本の中、観智院本（国宝）が唯一の完本である。現在は天理大学附属天理図書館に所蔵されている。

2. 2 HDIC について

平安時代漢字字書総合データベース（Integrated Database of Hanzi Dictionaries in Early Japan, HDIC と略す）は日本の平安時代に撰述された漢字字書をデータベース化し、インターネットで公開するプロジェクトである。収録する字書は高山寺本『篆隸万象名義』（略称 KTB）、天治本『新撰字鏡』（TSJ）、図書寮本『類聚名義抄』（ZRM）、観智院本『類聚名義抄』（KRM）である[2]。

3. 字形情報利用の制限

3. 1 符号化による制限

HDIC データベースは Unicode によって符号化されている。Unicode の収録字数は年々増加しているが、その収録方針によって、古辞書に収録した異体字が少なからず未登録のものがある。独自包摂基準を設定し、翻刻文字は原本と全く同様にするには限界がある。さらに、データベースにアクセスする端末には最新の Unicode 漢字を表示させるフォントの実装も期待できず、古辞書データベースのインターフェースとして、画像で漢字を表示することが避けられない。

3. 2 著作権による制限

観智院本『類聚名義抄』は現在天理図書館に所蔵され、KRM での翻刻テキスト公開は同館から特別にご許可を賜り推進している。掲出字の原本画像を掲載・転載する際には、同館の要項に準じる[3]。HDIC プロジェクトとしての利用は、許可を得たが、二次利用に関して、利用者による再度申請が必要となる。

4. GlyphWiki によったの字形再現

4. 1 GlyphWiki について

GlyphWiki は 2007 年 10 月から公開されている漢字字形自由共有サイトである。GlyphWiki では、ユーザーが漢字字形を登録・編集することによって、インターネット上で漢字字形が共有可能となり、字形画像や外字フォントも同時に自動生成される。字形の編集には、専用エディタを採用し、平易に字形の骨格情報を操作できることから、これまでに大量の漢字・異体字の字形データが登録されている。

GlyphWiki は、用途が限定されない。よって、開発者はあるテキストデータに含まれる外字集合をまとめて登録し、GlyphWiki を通じて外字フォントを公開することが可能となる。

4. 2 字形データと Kage システム

GlyphWiki では、文字の情報しか登録できない。当サイトの字形専用エディタを利用して、字形を筆画ごとに直線、曲線などの線種情報と、線の座標情報、および「はね」「ねらい」「とめ」などの端形状情報の 3 種類情報を組み合わせで記録する[4]。GlyphWiki の同一開発者による Kage システムは、これらの情報を利用して、字形の画像とフォントを生成できる。

4. 3 作字プラットフォームの実現

作字プラットフォームは、HDIC の公開用インターフェース HDIC Viewer²から構築する。HDIC Viewer に格納している観智院本『類聚名義抄』テキストデータベースを用いて、ID と GlyphWiki に作成する字形³のリンクを生成して、所在・原本画像・作成した字形・注文などを一つのページにまとめる。

現在、HDIC Viewer の“/krm-glyphwiki”から部首の索引をアクセスできる（図 1）。

観智院本『類聚名義抄』			GlyphWiki画像作字用			部首索引		
巻	冊	字	巻	冊	字	巻	冊	字
1	1	人	5	41	水	8	81	神
1	2	フ	5	42	火	8	82	竹
1	3	電	5	43	土	8	83	刀
1	4	心	5	44	金	8	84	刀
1	5	木	5	45	石	8	85	目
1	6	草	5	46	虫	8	86	虫
1	7	鳥	5	47	魚	8	87	魚
1	8	獸	5	48	人	8	88	人

図 1 作字プラットフォーム部首索引（部分）

² <https://hdic2.let.hokudai.ac.jp>

³ GlyphWiki ではグリフと呼ぶ。

部首のリンクをクリックすると、該当部首の作字リストページを表示する(図2)。

No	Charge	Hkm-ID	Hkm	Glyph	Text
1	linchuj	Hkm-03038610	手	手	言守 字(山) 字ツカラ(LH) (山)
2	linchuj	Hkm-03038620	才	才	今更作此類字之生 ヒサク
3	linchuj	Hkm-03038630	春	春	古
4	linchuj	Hkm-03038641	手	手	タナス工(LH)
5	linchuj	Hkm-03038642	子	子	タナス工(LH)

図2 作字プラットフォームリストページ

作字済みの字形は原本画像の右に表示され、未完成の字形は赤色の“×”符号となる。その符号をクリックすると、作字担当者が事前に登録した GlyphWiki ユーザー名を用いて、字形を登録する予定である GlyphWiki ページへ移動できる。GlyphWiki の字形 URL に関して次の5.1で詳述する。

5. 再現した字形とその利用

5.1 再現した字形のアクセス

字形を GlyphWiki に登録する際、その名前は最初に決めなければならない。無秩序な登録を防ぐため、名前の登録には様々なルールがある[5]。また、GlyphWiki に登録した字形は誰でも修正・更新できることを基本とするが、確認・管理上の問題を避けるため、現段階では GlyphWiki に登録する字形は登録者しか修正・更新できない“ユーザー占有グリフ”としている。作字担当者が事前に各自 GlyphWiki のアカウントを登録して、作字担当者アカウントの“ユーザー占有グリフ”に字形を作成する。

各担当部分の字形の登録・作字を字形登録確認者の李媛が字形の確認を行った後、グリフ登録確認者のアカウント“hdic”のユーザー占有グリフに移行する。

GlyphWiki に登録する名前は hdic(ユーザー名)と, hkrm(観智院本『類聚名義抄』という集合名)とその所在情報の組み合わせで設定する。URL は“glyphwiki.org/wiki/”と字形の名前からなる。次は登録済みの字形 URL の例である。https://glyphwiki.org/wiki/hdic_hkrm-03038630

⁴ <http://www.chise.org/ids/index.ja.html>

⁵ KRM は Unicode によって符号化されたので、実際利用したファイルは IDS-UCS-Basic.txt, IDS-UCS-Ext-A.txt, IDS-UCS-Ext-B-1.txt, IDS-UCS-Ext-B-2.txt, ID

5.2 観智院本類聚名義抄データベースの画像代用

HDIC Viewer は著作権とデータ精度を考慮して、現在 KRM のインターフェースは公開版と内部版とに分けている。内部版では、ユーザー登録制を採用して、観智院本の掲出字の原本画像を表示することができるユーザーは研究グループのメンバーに限定。

一方、公開版では、今まで部分的なテキストしか公開されていなかった。しかし、GlyphWiki に登録した字形を導入することで、掲出字を原本画像の代わりとして表示することができるようになった(図3)。

順	所在(風間)	所在(天理)	部首	GlyphWiki	掲出字
1	仏下本帖39頁	仏巻301頁	手	駢	駢 / - (筆)

図3 HDIC Viewer の KRM 公開版検索結果

5.4 IDS を用いて作成した字形の検索

現時点まで、作成済みの字形は KRM のウェブインターフェースを用いて、翻刻された Unicode 漢字または所在情報で検索できる。文字同定や字形集成などの研究活動を行う際、同一漢字部品を含む漢字を特定する機能が求められる。つまり、字形で字形を検索することが必要となる。

しかし、現時点では KRM の字形データがまだ完成度が低く、カバーできる掲出字数が少ないため、符号レベルで漢字の部品検索する機能を実装することが現実である。

部品検索の実現は以下のステップで行う。

①部品データの準備:CHISE 漢字構造情報データベース⁴を用いて⁵、漢字一部品から部品一漢字索引を作成する(図4)。このようなデータ用意すれば、インターフェース側はより少ない計算で漢字構成した最小部品を検索すること

S-UCS-Ext-B-3.txt, IDS-UCS-Ext-B-4.txt, IDS-UCS-Ext-B-5.txt, IDS-UCS-Ext-B-6.txt, IDS-UCS-Ext-C.txt, ID S-UCS-Ext-D.txt, IDS-UCS-Ext-E.txt, IDS-UCS-Ext-F.txt, IDS-UCS-Ext-G.txt である。

ができる。

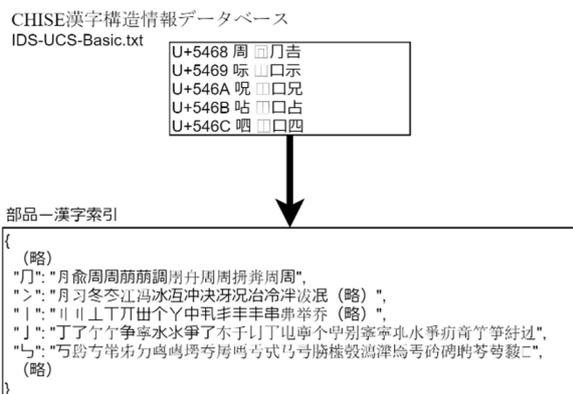


図 4 部品一漢字索引の生成

②インターフェース側の検索機能: オープンソース部品筆画漢字検索システム「零時字引」⁶を参考して, 部品と残りの筆画数で漢字を検索する機能を実装した. 検索結果を KRM の Entry フィールドと照合して, 掲出字のみのデータを表示する.

③字形の表示: すでに GlyphWiki に字形が登録された掲出字は, 画像の字形を表示する. 実装したシステムは図 5 で示す.

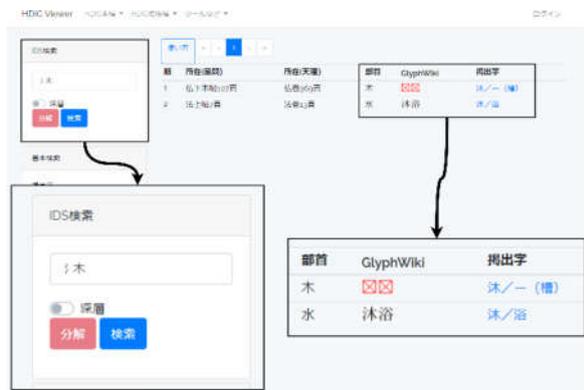


図 5 部品字形検索画面

KRM は Unicode を用いて符号化して, 漢字構成の IDS⁷データは入手しやすいが, 符号と表示するフォントの対応関係は一对多のため, 求める字形と異なる検索結果が出てくる可能性がある. このような字体と字形との対応に関して, 5. 4 節で詳しく述べる.

5. 3 字形データ

先述のように, GlyphWiki に保存している字形の情報は Kage システムを用いて記録される (図 6). Kage データは筆画の線の種類, 位置, 端形状からなっており, さらに別の掲出字に再利用することができる. 単純な輪郭データと比べると, より筆画の本質的な情報を使うため, データサイズは小さい[6].

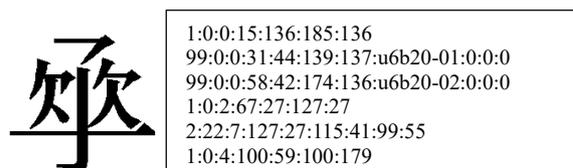


図 6 GlyphWiki の「手」の異体字とその Kage データ

写本字形を Kage データで記録することによって, IDS 方式より厳密に字形が相違する部位の標出や筆画単位で諸本間の比較など, 写本文字の情報学的分析が可能になる.

5. 4 字形情報と漢字の包摂粒度

IDS 情報は符号化された漢字の構造と利用された部品からなったものである. IDS の差分のより, 漢字符号の構成相違を比較することができるが, テキストデータ化された漢字文献を利用する際に, 漢字の包摂基準を考えないといけない. 特に, データ中の符号を字形研究の根拠として扱う場合, どこまで包摂された情報, つまり, 漢字の包摂粒度は重要な指標である. 包摂粒度について, 守岡知彦氏は以下のような指摘がある.

漢字の包摂粒度を考える場合, 常用漢字表で用いられている『字種』-『字体』-『字形』という概念と, UCS 等の符号化文字集合で用いられている『抽象文字』-『グリフ』-『グリフイメージ』という概念等があるが, これらをざっくりとまとめると, 大まかに言って, 『字種』-『抽象文字』-『字体』≒『グリフ』-『字形』≒『グリフイメージ』という 4 階層の包摂粒度に整理することができる. [7]

KRM では, 掲出字を Entry (掲出字) と Entry_original (掲出字原本字形) との二つのフィールドに保存している. 前者は康熙字典体に

⁶ <https://github.com/g0v/z0y>

⁷ Ideographic Description Sequence, 漢字構成記述文字列.

包摂するのを原則として入力され、検索上に便利である。後者は作成者の判断によって、必要がある際に Unicode の範囲でなるべく原本に近い符号を選んだものである [8]。上記と 4 階層と照合すると、表 1 のように示される。

表 1 KRM の包摂粒度

字種	抽象文字	字体(グリフ)	字形(グリフイメージ)
Entry	Entr y	Entry または Entry_original	利用者端末のフォント

KRM の利用者端末における漢字環境は複雑であり、作成者が作成時に意図した字形のまま表示されるとは限らない。そのため、Unicode に登録された字形であっても、GlyphWiki を用いてグリフイメージを用意する。

5. 5 作成字形と Unicode の関係

前述のように、Kage データは漢字の筆画を記述することによって字形を保存することである。また、複数の筆画をグループとして、別の字や部品を引用することができる。図 6 の右にある Kage データの 2 行目、99 から始まるデータを例として挙げられる。

なお、GlyphWiki で字形を作成する際、既存する字形を引用することは強制的ではない。引用せず、一つの筆画を一行のデータにしても、画像的に字形の表現ができる。引用した字形を筆画に分割することができるが、GlyphWiki 上の関連字と名前を失うことになる。したがって、本研究における字形作成では、既存字形を積極的に引用することを方針にした。既存字形がない場合のみ、筆画情報で登録する。

現時点において、GlyphWiki の「hdic」ユーザーが占有する 2,100 個の字形を分析する。KRM に翻刻された Entry 欄⁸に格納する掲出字との相違部分、およびよく出現するが引用できる字形がない部品を算出してみる (表 2)。

表 2 作成字形と Unicode の関係

分類	字数	
Entry のまま引用	870	
新作	部分 Entry の部品引用	765
	Entry の部品引用せず	75
対象外	IDS	359
	代用符号	31
合計	2,100	

⁸ Entry_orignal が IDS で記述される場合が多いため、今回は比較対象にしない。

翻刻した掲出字の Entry と GlyphWiki での作字と比較して、部品レベルで字形差があるものはほぼ半分 (対象外以外) を占めていることが分かる。

Entry のまま引用したのは 5. 4 節で述べたグリフイメージとしての字形である。例を挙げると、観智院本『類聚名義抄』の仏下本 38 頁にある“手: 音守, テ, テツカラ”の掲出字「手」は、GlyphWiki に登録されたグリフデータが“99:0:0:0:200:200:u624b-j”⁹である (図 7)。つまり、Unicode に登録した日本ソースの“手” — “u624b-j” という名前の字形のみを引用したことである。



図 7 GlyphWiki のグリフデータ形式

部分 Entry の部品引用は、部品は部分に Entry の掲出字と同様である新作字形である。仏下本 126 頁にある“齧: 音葵(蔡), 芥”の掲出字“齧”は例として挙げられる。作成された字形は、“葵”となって、グリフデータは次に示す。

```
99:0:0:8:75:187:194:u708e-j:0:0:0
99:0:0:-1:52:209:200:u5196-03:0:0:0
99:0:0:7:1:195:150:u6797-03:0:0:0
```

“葵”は u708e の“炎”, u5196 の“一”, u6797 の“林”の三つの字形からなった。翻刻された“齧”と比べると、“一”と“口”の相違がみられる。このような字形は、合計 1,760 回部分 Entry の部品字形を引用した。一字形は平均約 2.3 回を引用する。

字形のグリフデータを徐々に蓄積すれば、翻刻された漢字と原本字形との差分を明らかにすることができる。さらに諸本のグリフデータを揃えば、諸本間の字形相違を情報学手法で観察することが期待できる。

⁹ https://glyphwiki.org/wiki/hdic_hkrm-03038610

6. おわりに

以上, GlyphWiki を利用して, 観智院本『類聚名義抄』の掲出字を原本字形に近い形式に再現することを試みた. 現在までの作業結果は4部首(人部・水部・手部・木部)となる. 将来的には, 登録した字形のIDSを追加して, 字形の構造と利用した漢字部品の情報を連関させることを考えている. これらの字形データを蓄積することで, 漢字規格への符号登録や情報学的字体分析の一助となることを期待している.

付記

本研究は JSPS 科研費 19J11614, 19H00526 の助成を受けたものである.

参考文献

- [1] 李媛, 劉冠偉, 池田証壽. 観智院本類聚名義抄に関する作字作業をめぐって. 東洋学へのコンピュータ利用第32回研究セミナー. 2020, p.27-48.
- [2] 池田証壽, 李媛, 申雄哲, 賈智, 齋木正直. 平安時代漢字字書のリレーションシップ. 日本語の研究. 2016, Vol. 12, No. 2, p. 68-75.
- [3] “天理図書館 所蔵資料掲載・転載要項”. <https://www.tcl.gr.jp/usage-guidance/copy/copy-essential-point/>. (参照 2020-09-06).
- [4] 上地宏一. GlyphWiki: 漢字字形自由共有サイト. 学芸国語国文学. 2018, Vol. 48, p. 181-191.
- [5] “GlyphWiki: グリフを登録する”. <https://glyphwiki.org/wiki/GlyphWiki:グリフを登録する>, (参照 2020-09-07).
- [6] 上地宏一. 漢字グリフ管理 Wiki システム (GlyphWiki) の構築. じんもんこん 2007 論文集. 2007, p. 237-244.
- [7] 守岡知彦. CHISE における漢字字体・字形粒度の整理規準について. 東洋学へのコンピュータ利用 第26回研究セミナー. 2015, p. 153-190.
- [8] 池田証壽, 劉冠偉, 鄭門鎬, 張馨方, 李媛. 観智院本『類聚名義抄』全文テキストデータベース: その構築方法と掲出項目数等の計量 (小特集 類聚名義抄). 訓点語と訓点資料. 2020, Vol. 144, p. 129-105.