Vol.2012-CH-94 No.8
2012/5/26

漢字字形データベースと文字オントロジーのデータ統合の可能性について

守岡 知彦[1,a)]

概要：漢字字形共有サービス GlyphWiki と文字オントロジー共有サービス CHISE-wiki を利用者から見て一体のシステムとして運用できるように、主に、グリフ名/素性名の対応関係に着目して議論する。

キーワード：漢字，グリフ，字形，文字オントロジー

Possibilities of integration between a glyph-image database for Kanji characters and a character ontology

Morioka Tomohiko[1,a)]

Abstract:
This paper discusses how to integrate GlyphWiki and CHISE-Wiki, while GlyphWiki is a glyph-image database sharing system for Kanji characters and CHISE-Wiki is a character ontology sharing system. In particular, it focuses mapping between glyph-name of GlyphWiki and character feature of CHISE to manage GlyphWiki and CHISE-Wiki as one system in a view of users.

Keywords: Kanji, CJK Ideographs, glyph, glyph-image, character ontology

1. はじめに

漢字はその文字数の多さとその長期にわたる歴史的変遷の結果等により、時代や地域によって使われる字種が変化したり、形が変化したり、同形の文字の音義が変化したり、規範が変化したため、多様な異体字・類字関係が存在したり、同一性やカテゴライズ方法が簡単に定義できなくなったり、現代では廃れてしまった文字が出土文献等から新たに発見されたりするため、UCS [1] 等の標準的な符号化文字集合には存在しない文字を外字として扱う必要性がなかなかなくならないといえる。[2] しかしながら、外字はインターネットでの情報交換にとって問題があり、標準的な符号化文字集合に収録することで外字の使用を排除しようとする努力が続けられて来た。しかしながら、漢字の性質を考えれば、ある有限の文字の集合に存在しない字は常に現れうるといえ、符号化できない字が見つかった時に、その字の収録が済むまでその文字を使用できないとするのは不便であるし、また、全ての字を収録しなければならないとすれば、通常のテキストではほとんど使われないであろう文字まで収録しなければならないことになり、工業標準の経済性という観点で問題があるかも知れない。いずれにせよ、最終的に UCS 等の標準的な符号化文字集合に収録するとしても、漢字を自由に定義しインターネット上で交換可能にするための仕組みは必要であるといえる。

標準的な符号化文字集合に収録されていない文字を既に収録されている文字と同様に電子テキスト中で使用し、インターネット上で交換可能にするためには、単に外字の字形を画像情報として定義するだけでは不十分であり、標準化された符号化文字と同様に、外字に関するさまざまな属性や知識、すなわち、計算機においてさまざまな処理をする上で必要なさまざまな情報やその外字に関する存在論

[1] 京都大学人文科学研究所
Institute for Research in Humanities, Kyoto Uniersity
[a)] tomo@zinbun.kyoto-u.ac.jp

的な情報（どういう文字なのか？（どの範囲のものを包摂しているのか？）といった意味論的な情報）を機械可読な形で記述する必要があるといえる。著者らが提案しているChaon モデルやその実装である CHISE [3] [4] は文字のメタデータやオントロジーを用いることで、特定の符号化文字集合に依存しない文字処理技術を実現することにより、この問題を原理的に解決するための枠組を提供するものである。

一方、電子テキストを人間が読み書きするためには文字を表示する必要があり、そのためには字形をインターネット上で交換するための仕組みも必要である。上地宏一氏が開発した GlyphWiki [5] はインターネット上で漢字の字形を簡単に作成し共有することを可能とする極めて野心的な試みのひとつである。GlyphWiki では Wiki 的なアプローチを用い、既存の漢字部品を流用することで簡単に作字可能なユーザーインターフェースを提供するとともに、含んでいる部品オブジェクトとの関係や異体字関係といった字形に関するメタデータを管理することで字形情報の検索可能性を高めており、複数人で協力して大量の漢字字形を開発し、それを共有するための基盤を提供している。この仕組みにより、実際に有志の協力によって UCS の全文字を網羅するフォント（花園明朝）の開発に成功し、今も成長し続けている。

GlyphWiki は基本的に漢字字形を対象としたシステムであるので、そこで扱われる情報は基本的に字形やグリフに関するものに限られる。例えば、抽象文字といった単位は本質的に存在せず、[*1] 文字に関するメタデータは CHISE や Unihan データベース [6] といった外部の情報源に委ねている。また、現在の所、抽象文字としては UCS の符号化文字のみが扱われており、CHISE [*2] に対しても UCS の抽象文字に相当する文字オブジェクトへのリンクしか張られておらず、それ以外のさまざまな例示字形オブジェクトへのリンクは張られていない。

他方、CHISE の文字オントロジーをインターネット上から利用するためのものとして著者は CHISE-Wiki [7] を開発している。これは CHISE の文字オントロジーを Wiki のように Web ベースで閲覧したり編集したりするためのサービスであり、構造データのための Wiki である EgT[8] を用いて実現している。CHISE-Wiki はそれ単体では漢字字形を編集するための仕組みを用意していないが、対応する GlyphWiki の頁が機械的に決定できる場合、CHISE-Wiki の頁中に GlyphWiki の字形を表示するとともに GlyphWiki の該当頁へのリンクを張ることですぐに GlyphWiki で編集できるようにしている。

もし、GlyphWiki と CHISE-Wiki がより密接に連係し、利用者からみて一体のシステムとして利用可能な仕組みを実現すれば、本当の意味で漢字を自由に定義・交換可能な環境を実現することができると考えられる。このためには、

(1) GlyphWiki の任意のグリフに対して CHISE のオブジェクトが一意に対応し、実際に GlyphWiki のグリフの頁から CHISE-Wiki の該当頁にリンクが張られること
(2) CHISE-Wiki の任意の漢字字形オブジェクトに対して GlyphWiki のグリフが一意に対応し、実際に CHISE-Wiki の漢字字形の頁から GlyphWiki の該当頁にリンクが張られること
(3) 新規頁を作ることができ、一方にしか存在しないオブジェクト（頁）からのリンクを開いた時に、該当する新規頁が作成できること

を実現する必要があると考えられる。

ここでは、こうした要求を実現するために必要となる GlyphWiki のグリフと CHISE のオブジェクトの（機械的に実行可能な）一意な対応を実現する上で必要となる事項を明らかにするために、まず、GlyphWiki と CHISE の文字オントロジーにおける漢字や漢字部品の管理・運用の現状について簡単にまとめるとともに、両者の対応可能性について議論する。特に GlyphWiki のグリフの命名ガイドラインと CHISE の素性名の命名規則やオブジェクト間の関係の記述法に着目して、それらの仕様が示す形式やその記述力、曖昧性や、両者の対応可能性、現状では対応しない部分といった問題点について考察するとともに、対応させるために必要な事項に関して議論する。

2. GlyphWiki

GlyphWiki は上地宏一氏が開発した Web ベースの漢字字形共有システムである。GlyphWiki は漢字字形（グリフ）を共同編集するための Wiki とし てデザインされており、Web ブラウザーを使って GlyphWiki のサイトにアクセスすることで、誰でも自由にグリフを新たに作成したり、既に作成されているグリフを修正したり、そうして作成・編集した結果を登録して、インターネット上で共有することができる。

2.1 グリフの命名ガイドライン

GlyphWiki では各グリフは固有の名前によって管理されている。[*3] GlyphWiki は、システム的には、グリフに対して利用者が自由に名前を付けることを許容しているが、符号化文字集合や標準的なグリフセット、辞書等の例示字

[*1] 一見、符号化文字集合のコードポイントを代表するようなオブジェクトのように見えたとしても、それはあくまで例示字形等の『グリフ』（字形）を表現したものであり、抽象文字を表現したものではないといえる。また、関連グリフもあくまでグリフ間の対等な関係を表現しており、抽象⇔具象関係といったグリフ以外の粒度のオブジェクトとの包含関係のようなものが用意されている訳ではない。
[*2] 現在は CHISE-Wiki に対するリンクとなっている。
[*3] グリフ名にはエイリアスを付けることもできる。

形やある程度共有されている外字集合等のグリフ名に関してはその命名に関してガイドラインを設けている。また、部品の変形や変種、異体字等を示す修飾子のようなものも規定されている。

2.2 形式

GlyphWiki の命名ガイドラインは次のような ABNF 表記で表現できる：

```
Glyph-Name = Component / Squared-Words
Component = Component-Base [ Variant-Specifier ]
Component-Base = Modified-Component / IDS
Modified-Component = UCS-Component
                   / Other-Component
UCS-Component = UCS-Char [ UCS-Glyph-Modifier ]
UCS-Char = “u” 4*5 ( DIGIT / “a” / “b” / “c”
                   / “d” / “e” / “f” )
UCS-Glyph-Modifier
    = “-” UCS-Source-Specifier Component-Modifier
UCS-Source-Specifier = “” / “g” / “t” / “j” / “k”
                     / “v” / “h” / “kp” / “u”
                     / “m” / “us” / “i” / “ja”
                     / “js” / “jv”
Other-Component = Other-Component-Base
                [ “-” Component-Modifier ]
Component-Modifier = 2DIGIT
Other-Component-Base = IVS / Other-Coded-Glyph
Variant-Specifier
    = “-” ( “var” / “itaiji” ) “-” 3DIGIT
IDS = ( IDC2 “-” IDSs “-” IDSs )
      / (IDC3 “-” IDSs “-” IDSs “-” IDSs)
IDC2 = “u2ff0” / “u2ff1”
       / “u2ff4” / “u2ff5” / “u2ff6” / “u2ff7”
       / “u2ff8” / “u2ff9” / “u2ffa” / “u2ffb”
IDC3 = “u2ff2” / “u2ff3”
IDSs = UCS-Char / IDS
Squared-Words = “kumimoji-” 2*( UCS-Char )
```

すなわち、組文字 (Squared-Words) を別にすれば、GlyphWiki のグリフ名は、基本的に、(符号化文字集合の種類とそこでのコードポイント等で (その例示字形によって) 示される) ベースとなるグリフ名に、その部品配置による変形 (『偏化変形』) の種類を示す修飾子 (Component-Modifier) と異体字修飾子 (Variant-Specifier) を付けた形で構成されるといえる。Component-Modifier と Variant-Specifier はともに省略可であるが、その順番は先に Component-Modifier が来て最後に Variant-Specifier 来る形式となっており、修飾子のネストは認められていないと考えられる。しかしながら、ガイドラインの記述には曖昧性があり、別の解釈が成立する余地があり、実際、“u241fe-itaiji-001-03”, “u2696f-itaiji-001-03”, “u5de5-itaiji-001-02”, “u826f-itaiji-001-02” という例がある。ただ、これは少数例であり、大部分はここで示した形式に則っているようである。

code	意味
01	左右結合の左、左中右結合の左、中
02	左右結合の右、左中右結合の右
03	上下結合の上、上中下結合の上、中
04	上下結合の下、上中下結合の下
05	囲い結合の外
06	囲い結合の中
07	位置の指定はないが単独字ではなく部品として利用
08	縦長部品として利用（01、02 の共通部品に相当）
09	横長部品として利用（03、04 の共通部品に相当）
10	囲い外部品で中の密度が通常より高いもの
11	囲い外部品で中の密度が通常より低いもの
14	上下結合の下、上中下結合の下で、三角屋根の形状のもの
15	05 以外の同じ UCS コードポイントに対する囲い外部品

表 1　偏化変形部品用接尾コード

2.3 Component-Modifier

漢字部品は、偏（左右に配置する場合の左）や旁（同右）、冠（上下に配置する場合の上）や脚（同下）のように、配置する場所によって変形する場合がある。このような部品配置による変形のことを GlyphWiki では『偏化変形』と呼び、このの種類を示す修飾子 (Component-Modifier) に対して 10 進 2 桁の接尾コードを振っている（表 1)。

2.4 UCS-Source-Specifier

UCS (ISO/IEC 10646, Unicode) の符号位置に対応するグリフ（例示字形）の名前には、複数欄表記のどの欄の例示字形かを表 2 に示す接尾コード (UCS-Source-Specifier) を利用して明示する必要がある。

code	意味
なし	漸次廃止
g	G（中国）ソース
t	T（台湾）ソース
j	J（日本）ソース
k	K（韓国）ソース
v	V（ベトナム）ソース
h	H（香港）ソース
kp	KP（北朝鮮）ソース
u	U ソース（ISO 規格における U ソース）
m	M（マカオ）ソース
us	The Unicode Standard の字形
i	ISO 規格で Ext.B などの一欄表記となっているもの
ja	Ext.A の JA ソース
js	補助漢字
jv	仮想 J ソース

表 2　UCS のソース指定のための接尾コード

歴史的事情から、この接尾コードが省略されている例が多数存在するが、Component-Modifier（偏化変形修飾子）が指定された場合は “j” ないしは “jv” が指定されているものと看做すことになっているようである。[*4]

“u” はもともと「The Unicode Standard の字形」としていたものであるが、『The Unicode Standard の（一欄表記の場合の）字形』には “us” を用い、“u” は ISO/IEC 10646 における U ソースを示すものとなっている。

“jv” の「仮想 J ソース」というのは、J ソースがないものに対して、他の J ソースのグリフと混ぜても違和感がないようにデザインされたものを指す。具体的には、

- J ソースの漢字と混ぜて違和感のないもの
- 具体的には平成明朝体字体に沿うもの
- 筆画の接続の有無など細かい点は強制的な統一を図らない（議論の余地あり）

ということが規定されている。

2.5 Variant-Specifier

異体字に相当するグリフに対しては、ベースとなる漢字（部品）グリフの名前の後に「-var-*ddd*」もしくは「-itaiji-*ddd*」という異体字指示のための修飾子を付けることになっている。ここで、*ddd* は 10 進 3 桁の番号で、この番号は各符号位置ごとに「001」から順番につけるものとなっている。なおグリフの削除などで欠番が生じた場合には、その番号は廃止とし、再利用しないことになっている。また、「-var-*ddd*」と「-itaiji-*ddd*」は独立した名前空間になっており、同じ番号が振られても両者は関係づけられない。

「-var-*ddd*」と「-itaiji-*ddd*」の区別は、「ISO/IEC 10646 でユニフィケーション対象となっている差異」に相当するか否かを示している。すなわち、ユニファイされうる場合は「-var-*ddd*」を用い、ユニファイされないものには「-itaiji-*ddd*」を用いる訳である。[*5]

2.6 AdobeJapan 1 のグリフ

AdobeJapan 1 のグリフの場合、

CID 番号による表現　aj1-*ddddd*
（*ddddd* は 10 進 5 桁の CID 番号；例：aj1-07765）

IVS による表現　<基底文字>-<Variant Selector>
（基底文字と Variant Selector は UCS-Char 形式（“u*hhhh*(*h*)”）で表記；例：u90a3-ue0101）

の 2 種類を許容しており、特にどちらかに統一することを要求していない。

[*4] ガイドライン上は、“j” が指定されているものとみなすとあるが、実際には J 欄が存在しない文字も存在するので、その場合は “jv” が指定されていると看做さざるを得ないと思われる。
[*5] ユニファイされない異体字は将来 UCS において別のコードポイントが振られるかも知れない。

2.7 IDS

IDS (Ideographic Description Sequence) に基づき部品の組合せ（漢字構造情報）によってグリフを表現する記法も存在する。この記法では、オペレーターである IDC (Ideographic Description Characters) や各部品は UCS-Char 形式（“u*hhhh*(*h*)”）で表記され、各部品に対して UCS ソース指示子 (UCS-Source-Specifier) や偏化変形修飾子 (Component-Modifier), 異体字修飾子 (Variant-Specifier) を付けることを認めていない。

表現可能な例　u2ff0-u53e3-u6606

表現不可能な例　u2ff0-u53e3-01-u6606-02

ただ、IDS 全体に対して異体字修飾子を付けることは認められている。よって、もし、IDS を同じくする複数の異なるグリフを表現したい場合、IDS 全体に対して通番の異体字修飾子を付けることになる。

例　u2ff0-u53e3-u6606-itaiji-001 は表現可能

2.8 Other-Coded-Glyph

UCS 以外の幾つかの符号化文字集合やグリフセット、辞書等の例示字形に対して、命名規則が予約されている（表 3）。

3. CHISE

CHISE は著者らが開発している文字オントロジーに基づく文字処理技術で、実際に処理系と文字オントロジーやデータベースを公開している。CHISE は Chaon モデルと呼ぶ『確定記述の束』として文字を指示する手法に基づき文字を表現しており、各文字は素性名とその値からなる素性対の集合によって表現されている。

実際に CHISE の文字オントロジーに収録されている文字オブジェクトには、UCS 等の抽象文字に相当するものの他に、抽象字体レベル、抽象字形レベル、例示字形、複数の抽象文字を包摂する超抽象文字といったさまざまな粒度のものが含まれる。[9] よって、実際に GlyphWiki に対応しうるのは例示字形オブジェクト（あるいは、抽象字形レベルのオブジェクト）になると考えられる。

3.1 例示字形オブジェクト

CHISE の文字オントロジーに収録されている字形レベルのオブジェクトの多くはなんらかの符号化文字集合やグリフセット、辞書等に対応するものである。特に、このようななんらかの文字（グリフ）セットの例示字形を表現するオブジェクトを例示字形オブジェクトと呼ぶ。

例示字形オブジェクトは、1 つ以上のソースを持ち、そのソースを示す素性名とそのソースにおける番号からなる素性対を持つ。ここで、この例示字形オブジェクトのソースを示す素性名は『ID 素性』の一種であり、各例示字形

表記	意味
toki-*hhhhhhhh*	登記統一文字番号
koseki-*dddddd*	戸籍統一文字番号
juki-*hhhh*	住基ネット統一文字コード
gt-*ddddd*	GT コード
gt-k*ddddd*	GT-k コード
tron-*d-hhhh*	TRON コード
cdp-*hhhh*	CDP 外字
cb*ddddd*	CBETA
j78-*hhhh*	JIS X 0208:1978
j83-*hhhh*	JIS X 0208:1983
j90-*hhhh*	JIS X 0208:1990
jsp-*hhhh*	JIS X 0212:1990
jx1-2000-*hhhh*	JIS X 0213:2000 第 1 面
jx1-2004-*hhhh*	JIS X 0213:2004 第 1 面
jx2-*hhhh*	JIS X 0213 第 2 面
k0-*hhhh*	KS X 1001
c1-*hhhh*	CNS 11643 第 1 面
c2-*hhhh*	CNS 11643 第 2 面
c3-*hhhh*	CNS 11643 第 3 面
c4-*hhhh*	CNS 11643 第 4 面
c5-*hhhh*	CNS 11643 第 5 面
c6-*hhhh*	CNS 11643 第 6 面
c7-*hhhh*	CNS 11643 第 7 面
cf-*hhhh*	CNS 11643 第 15 面
b-*hhhh*	Big5 コード
jc3-*hhhh*	JEF-CHINA3 コード
dkw-*ddddd*	諸橋轍次『大漢和辞典』番号
dkw-*ddddd*d	同（ダッシュ付き）
dkw-*ddddd*dd	同（2 点ダッシュ付き）
dkw-h*dddd*	同（補巻）
kx-*ppppcc*	康熙字典（同文書局影印本）
waseikanji-no-jiten-*dddd*	和製漢字の辞典
kokuji-no-jiten-*dddd*	国字の字典
nihonjin-no-tsukutta-kanji-*ddd*	日本人の作った漢字
zihai-*ppppcc*	中華字海

但し、*pppp* は 10 進 4 桁のページ番号を表し、
cc は 10 進 2 桁のページ内番号を表す。

表 3 その他の文字コード、字典番号など

オブジェクトは必ず固有の番号を素性値として持つ（つまり、その素性値に対して同じ番号の異なるオブジェクトが存在しない；素性値を使って逆引可能であることが保証される）。

CHISE の文字オントロジーでは、このような例示字形オブジェクトのソースとなる ID 素性名（例示字形 ID 素性名）を “=*foo*” というシンボルで表すことになっている（表 4, 5, 6, 7, 8, 9, 10）。この素性名はソース毎に固有のシンボルが割り当てられている。なお、今の所、ID 素性の値は整数に限定されている。[*6]

これらの例示字形 ID 素性名の中には

(1) GlyphWiki と 1 対 1 対応するもの

[*6] Concord/EgT ではこの制限はない。

素性名	内容
=jis-x0208	JIS X0208（共通部分）
=jis-x0208@1978	JIS X 0208:1978（共通部分）
=jis-x0208@1978/1pr	JIS X 0208:1978（第 1 刷）
=jis-x0208@1978/-4pr	JIS X 0208:1978（第 1〜3 刷）
=jis-x0208@1978/-4X	JIS X 0208:1978（注 1）
=jis-x0208@1978/1er-pr	JIS X 0208:1978（注 2）
=jis-x0208@1978/2-pr	JIS X 0208:1978（第 2 刷以降）
=jis-x0208@1978/4er	JIS X 0208:1978（注 3）
=jis-x0208@1978/4-pr	JIS X 0208:1978（第 4 刷以降）
=jis-x0208@1978/5pr	JIS X 0208:1978（第 5 刷）
=jis-x0208@1983	JIS X0208:1983
=jis-x0208@1990	JIS X0208:1990
=jis-x0212	JIS X0212
=jis-x0213-1	JIS X 0213 第 1 面（共通部分）
=jis-x0213-1@2000	JIS X0213:2000 第 1 面
=jis-x0213-1@2004	JIS X0213:2004 第 1 面
=jis-x0213-2	JIS X0213 第 2 面

注 1：JIS X 0208 1978 年版 第 4 刷より前の規格票の字形索引に用いられ、第 4 刷附属の正誤表で置き換えられた字形
注 2：1978 年 11 月の正誤表で置き換えられた字形
注 3：第 4 刷附属の正誤表で置き換えが指示された字形

表 4 漢字関連の例示字形素性名 (1) JIS 関連

素性名	内容
=gb2312	GB2312
=ks-x1001	KS X1001
=iso-ir165	ISO-IR-165 (CCITT Extended GB)
=cns11643-1	CNS 11643 Plane 1
=cns11643-2	CNS 11643 Plane 2
=cns11643-3	CNS 11643 Plane 3
=cns11643-4	CNS 11643 Plane 4
=cns11643-5	CNS 11643 Plane 5
=cns11643-6	CNS 11643 Plane 6
=cns11643-7	CNS 11643 Plane 7

表 5 漢字関連の例示字形素性名 (2) JIS 以外の ISO-IR 関連

素性名	内容
=gb12345	GB 12345-1990
=big5	Big5
=big5-eten	Big5 ETEN
=adobe-japan1	Adobe-Japan1
=adobe-japan1-0	Adobe-Japan1-0
=adobe-japan1-1	Adobe-Japan1-1
=adobe-japan1-2	Adobe-Japan1-2
=adobe-japan1-3	Adobe-Japan1-3
=adobe-japan1-4	Adobe-Japan1-4
=adobe-japan1-5	Adobe-Japan1-5
=adobe-japan1-6	Adobe-Japan1-6

表 6 漢字関連の例示字形素性名 (3) 他の標準的符号

(2) 概ね対応するが GlyphWiki には存在しない区別を行っているもの

(3) GlyphWiki に対応するものがないもの

が存在している。

素性名	内容
=hanyo-denshi/ja	JA (JIS X0208)
=hanyo-denshi/jb	JB (JIS X0212)
=hanyo-denshi/jc	JC (JIS X0213:2000 Plane 1)
=hanyo-denshi/jd	JD (JIS X0213:2000 Plane 1)
=hanyo-denshi/ft	FT (FDPC 追加)
=hanyo-denshi/ia	IA
=hanyo-denshi/ib	IB
=hanyo-denshi/hg	HG（表外漢字表）
=hanyo-denshi/ip	IP (for IPA)
=hanyo-denshi/jt	JT（住基統一文字）
=hanyo-denshi/ks	KS（戸籍統一文字）

表 **7** 漢字関連の例示字形素性名 (4) 汎用電子関連

素性名	内容
=daikanwa	大漢和辭典（共通部分）
=daikanwa@rev1	大漢和辭典（修訂版）
=daikanwa@rev2	大漢和辭典（修訂第２版）
=daikanwa/+p	大漢和辭典（*ddddd'*）
=daikanwa/+2p	大漢和辭典（*ddddd''*）
=daikanwa/ho	大漢和辭典（補巻）
=shinjigen	角川新字源（共通部分）
=shinjigen@1ed	角川新字源（初版）
=shinjigen@1ed/24pr	角川新字源（初版 24 刷）
=shinjigen@rev	角川新字源（改訂版）
=shinjigen/+p@rev	角川新字源（改訂版；*dddd'*）

表 **8** 漢字関連の例示字形素性名 (5) 辞書類

素性名	内容
=big5-cdp	Big5 + CDP 外字
=gt	GT 2000
=gt-k	GT 部品集合
=hanziku-1	漢字庫（疑似 Big5 符号化）第 1 面
=hanziku-2	漢字庫（疑似 Big5 符号化）第 2 面
=hanziku-3	漢字庫（疑似 Big5 符号化）第 3 面
=hanziku-4	漢字庫（疑似 Big5 符号化）第 4 面
=hanziku-5	漢字庫（疑似 Big5 符号化）第 5 面
=hanziku-6	漢字庫（疑似 Big5 符号化）第 6 面
=hanziku-7	漢字庫（疑似 Big5 符号化）第 7 面
=hanziku-8	漢字庫（疑似 Big5 符号化）第 8 面
=hanziku-9	漢字庫（疑似 Big5 符号化）第 9 面
=hanziku-10	漢字庫（疑似 Big5 符号化）第 10 面
=hanziku-11	漢字庫（疑似 Big5 符号化）第 11 面
=hanziku-12	漢字庫（疑似 Big5 符号化）第 12 面
=cbeta	CBETA 外字
=zinbun-oracle	京大人文研所蔵甲骨文字
=jef-china3	JEF + CHINA3 外字
=ruimoku-v6	東洋学文献類目現行外字

表 **9** 漢字関連の例示字形素性名 (6) 外字集合等

(3) のケースに関しては GlyphWiki におけるグリフ名の付け方を考える必要があるといえる。

また、GlyphWiki に存在しているが、現在の所、CHISE には存在しないものがあり、GlyphWiki との対応を行うためにはこうしたものに対して CHISE の素性名を割り当てる必要があるといえる。

素性名	内容
=ucs@iso	ISO/IEC 10646 例示字形
=ucs@unicode	Unicode 例示字形
=ucs@gb	GB 例示字形
=ucs@cns	CNS 11643 例示字形
=ucs@jis	JIS X0208/0212/0213 例示字形
=ucs@jis/1990	JIS X 0208/0212:1990 例示字形
=ucs@jis/2000	JIS X 0213:2000 例示字形
=ucs@jis/2004	JIS X 0213:2004 例示字形
=ucs@JP	日本風デザイン
=ucs@JP/hanazono	花園明朝字形
=ucs@ks	KS 例示字形
=ucs@cns11643	www.cns11643.gov.tw の字形
=ucs@big5	Big5
=ucs@big5/cns11643	www.cns11643.gov.tw の Big5 字形

表 **10** 漢字関連の例示字形素性名 (7) UCS マッピング

また、CHISE の ID 素性名には継承機構があり、例えば、`=jis-x0208@1990` は `=jis-x0208` を継承している。これにより、もし版による差がない場合、共通部分を示す `=jis-x0208` によって記述されることになる（この場合、`=jis-x0208@1990` は `=jis-x0208` を継承しているので、単に `=jis-x0208` と書かれた箇所の素性値も自分のものとして扱うことになる)。

GlyphWiki にはそうした概念はなく、その代わりにエイリアスを用いて同一性を表現している。よって、CHISE から GlyphWiki にアクセスする場合、なんらかの優先度を設けて各インスタンス（すなわち、`=jis-x0208` や `=ucs@jis` のような共通部分を示す抽象的な素性ではなく、`=jis-x0208@1990` や `=ucs@jis/2004` のような具体的な版を示す素性を用いる必要があるといえる。逆に、GlyphWiki から CHISE にアクセスする場合、直接対応する具体的な素性でデコード処理を行って文字オブジェクトを得れば良い。

3.2 部品化変形

漢字は部品として使われる場合、単体の場合とは違った形に変形することがある。このような変形した部品に対し、CHISE では漢字間の異体字・類字関係の場合と同様に、関係素性を用いてその対応関係を記述する（表 11)。

素性名	意味
`<-formed`	～の異体
`<-same`	～に同じ
`<-identical`	～と同一
`<-original`	～の本字
`<-simplified`	～の略字
`<-vulgar`	～の俗字

表 **11** 部品化関係として使われている関係素性（使用例のあるもの）

サブドメイン名	意味	GlyphWiki
なし	部品化変形（一般）	07
connect-right	右に接続（偏）	01
connect-left	左に接続（旁）	02
connect-below	下に接続（冠）	03
connect-above	上に接続（脚）	04, 14
surround-from-above	上から囲む	05, 10, 11
surround-full	囲む	05, 10, 11
surround-from-below	下から囲む	05, 10, 11
surround-from-left	左から囲む	05, 10, 11
surround-from-upper-left	左上から囲む	05, 10, 11
surround-from-upper-right	右上から囲む	05, 10, 11
surround-from-lower-left	左下から囲む	15

表 12 部品化に関するサブドメイン
（上は使用例のあるもの。下はまだ使用例のないもの）

但し、部品化に関わる関係であることを示すために、component というドメインを付与する。また、部品は、偏(左右に配置する場合の左）や旁（同右)、冠（上下に配置する場合の上）や脚（同下）のように、配置する場所によって変形する場合があるが、こうした変形を示すために表 12 に示すサブドメインを用いる。

例　「人」<-formed@component/connect-right「亻」

また、ベースとなる部品として変形する前のオブジェクトを親とした親子関係を <-denotational 素性を使って記述する。

CHISE における部品化に関するドメインを GlyphWiki の『偏化変形』(表 1）と比較した場合、両者が 1 対 1 対応していないことが判る。GlyphWiki の方には『偏』(-01) と『旁』(-02) の共通部品に相当する『縦長部品化』(-08) というものや、『冠』(-03) と『脚』(-04) の共通部品に相当する『横長部品化』(-09) というものがある他、『囲い部品化（密度大)』(-10) や『囲い部品化（密度小)』(-11) といった密度に関する概念や、『三角屋根の形状の脚』(-14) や『その他の囲い部品化』(-15) [*7] といったものがあり、現状の CHISE のものよりも記述力が高くなっているといえる。これは GlyphWiki が実際に字形合成を行うからだと考えられる。一方、CHISE では囲む場合の変形に対して(IDC と同様な）囲み方の種類に応じた区別を設けている。

この問題を解決するためには、CHISE における囲み系ドメインを component/surround-from-above のような形から、component/surround/from-above のような形に変え、component/surround のような囲み方の種類に応じた区別を省略したドメインを認めることと、component/wide と component/tall といった GlyphWiki の -08 と -09 に対応するドメインを設けること、同様に -10 と -11 に対応する component/surround のサブドメインを設けること、そして、-14 に相当する component/connect-above のサブドメインを設けるのが良いと思われる。

また、現状の CHISE では関係素性を使う訳であるが、この場合、値にとるオブジェクトは複数存在し得るので、GlyphWiki から対応する CHISE のオブジェクトを得ようとする場合に曖昧性が生じるかも知れない。

[*7] 実際の用例をみる限り、これは『左下から囲む』に相当すると思われる。

4. 課題と対策

4.1 GlyphWiki におけるグリフ名の問題

4.1.1 グリフ名の形式化の問題

GlyphWiki のグリフ名は、現状では、一応のガイドラインはあるものの、それが十分に形式化されておらず曖昧性があるという問題があり、また、このガイドラインに基づかないグリフ名も付けることができるという問題もある。前者に関しては本稿で一応の形式化を行ってみたが、これがはたして妥当かという問題と、この形式に合致しない例をどうするかという問題が生じる。

また、このガイドラインが示す GlyphWiki のグリフ名の仕様ではグリフをベースとなる抽象部品に対して部品化(『偏化』）変形と異体字という 2 つの修飾子（UCS の場合にはソース指定子も）を指定することでさまざまな部品変種を記述可能にしていると考えられるが、この修飾子のネストを許していない（と考えられる）ために表現力が制限されている面があるかも知れない。しかしながら、無制限なネストを許すことにも問題があるといえる。こうしたことを鑑みれば、本稿での形式化に合致しない例を救済するためにも、2 段階までのネスト、すなわち、異体字修飾子に対する部品化修飾子の付加を認めるような拡張を行うのが良いかも知れない。

いずれにしても、GlyphWiki のグリフ名のガイドラインが要請する仕様を十分に形式化して曖昧性をなくす必要があると思われる。また、少なくとも公開を目的とするグリフの名前はガイドラインに沿ったものを強制するかガイドラインに則ったエイリアスが自動的に付くような仕組みがあると良いかも知れない。

4.1.2 IDS と部品の問題

GlyphWiki では文字グリフを部品として再利用するための仕組みを持っており、グリフ名のガイドラインもそうした発想に基づいて設計されていると思われる。しかしながら、IDS（あるいは組文字）による記述においては、部品として UCS のコードポイントしか表現できず、UCS にない部品や部品化・異体字修飾子を付けた異体部品を利用することができない。[*8] また、IVS のような UCS のコードポイントのシーケンスと IDS のような部品のシーケンスと修飾子の付与を示すセパレーターがともに “-” で示され、どこまでが部品なのかが判りにくいという問題があるように思われる。[*9] この問題を解決するには、IDS（と組文字）に

[*8] 実際には、UCS にない部品を用いた例は存在している。
[*9] おそらく、このために IDS で任意の部品を用いることができな

おける部品のセパレーターを “-” 以外に変えるなどして、修飾子のセパレーターと部品結合を示すセパレーターの区別が付くようにすることが考えられる。この場合、過去との互換性のため、新たに拡張した IDS には “ids_” のような接頭辞を付与すれば良いのではないかと思われる。これにより、例えば、

Glyph-Name
= Component / Ext-IDS / Squared-Word
Ext-IDS = “ids_” eIDSb
eIDSb = (IDC2 “_” eIDSs “_” eIDSs)
/ (IDC3 “_” eIDSs “_” eIDSs “_” eIDSs)
eIDSs = Component / eIDSb

という風にガイドラインの形式を拡張する訳である。

4.2 CHISE 側での対処

GlyphWiki のグリフに対して CHISE のオブジェクトを一意に対応させるためには、GlyphWiki にあって CHISE にないものをちゃんと CHISE の素性名として表現できるようにする必要があるといえる。

IDS や組文字による表現を除けば、GlyphWiki のグリフ名は（ガイドラインに則っている限り）ベースとなる符号化文字集合とそのコードポイントの対に対して修飾子を付けるという形になっているので、符号化文字集合と修飾子の組合せに対して CHISE の文字素性名が機械的に対応するようにすれば良いといえる。

一方、IDS や組文字、あるいは、ガイドラインに則っていないものにも対処したい場合、CHISE の ID 素性の値が自然数に限定されるという現状の制約を取り除き、値として文字列をとることを認めた上で、例えば `=glyph-id@glyphwiki` というような ID 素性名を設け、その値に GlyphWiki におけるグリフ名を入れるというような方法も考えられる。

ただ、GlyphWiki におけるエイリアスの存在や現行の CHISE の枠組との親和性を鑑みれば、なるべく CHISE の素性名に対応させる方法の方が望ましいと考えられる。よって、この両者を組み合わせるのが良いのではないかと思われる。

5. おわりに

漢字字形共有サービス GlyphWiki と 文字オントロジー共有サービス CHISE-wiki を利用者から見て一体のシステムとして運用できるように、主に、グリフ名/素性名の対応関係に着目して議論した。

GlyphWiki のグリフと CHISE の例示字形オブジェクトを一意に対応させるためには、GlyphWiki 側におけるグリフ名の命名規則の形式化を徹底することが望ましいと考えられる。そのために、本稿ではグリフ名の命名に関するガイドラインの形式化を試みたが、曖昧性の排除と実際の運用との親和性を鑑みれば、若干の拡張を行うことが望ましいかも知れない。また、CHISE 側では GlyphWiki のグリフ名をより柔軟に扱えるようにするための拡張を行うことが望ましいといえる。

また、絶えず修正可能な Wiki 的枠組においては、常に理想とする状態と現状の差が存在することや絶えず流動し続けるということを前提に考える必要があると思われる。この際、ある瞬間の状態を意味のある形で参照するための仕組みを GlyphWiki と CHISE のより密接な連係を実現することで整えて行くことも重要な課題のひとつであろう。

最後に、GlyphWiki と CHISE の連携に関してたびたび議論の機会を与えて頂いた上地宏一氏に感謝する。しかしながら、本稿における誤りは全て著者の責に帰すものであるのは言うまでもない。

いのだと思われる。

参考文献

[1] International Organization for Standardization (ISO): *Information technology — Universal Multiple-Octet Coded Character Set (UCS)* (2011). ISO/IEC 10646:2011.
[2] 守岡知彦：類目外字における “Old Hanzi”，東洋学へのコンピューター利用第 20 回研究セミナー，pp. 115–133 (2009).
[3] 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 (2006). 2006-CH-72.
[4] Morioka, T.: CHISE: Character Processing based on Character Ontology, *Large-scale Knowledge Resources (LKR2008)*, LNAI, No. 4938, pp. 148–162 (2008).
[5] 上地宏一：GlyphWiki, `http://glyphwiki.org/wiki/GlyphWiki`.
[6] : Unihan Database, `http://www.unicode.org/charts/unihan.html`.
[7] 守岡知彦：CHISE のセマンティック Wiki 化の試み，情処研報， Vol. 2010-CH-87, No. 8, pp. 1–8 (2010).
[8] 守岡知彦：Wiki 的手法に基づく構造化データの編集について，人文科学とコンピュータシンポジウム論文集—人文工学の可能性〜異分野融合による「実質化」の方法〜，情報処理学会シンポジウムシリーズ，Vol. 2010, No. 15, 情報処理学会，情報処理学会，pp. 33–40 (2010).
[9] 守岡知彦：CHISE に基づくグリフ・オントロジーの試み，人文科学とコンピュータシンポジウム論文集—デジタル・ヒューマニティーズの可能性，情報処理学会シンポジウムシリーズ，Vol. 2009, No. 16, 情報処理学会，情報処理学会，pp. 9–14 (2009).