多粒度漢字構造モデルに基づく字形整理の試み

— 漢字字体規範史データベースの CHISE への収録を通じて —

守岡 知彦 （京都大学）

CHISE 文字オントロジーで採用している多粒度漢字構造モデルは現在使われている漢字を整理する上では一定の成果をおさめているが、前近代の多彩な漢字字形を対象にした場合にどうなるかについては明らかでなかった。また、各包摂粒度の包摂範囲を合理的に規定するためには字体・字形用例の存在が重要であり、CHISE 文字オントロジーにグリフデータベースやグリフコーパスを統合することが望ましい。そこで、本研究では漢字を対象とした代表的グリフデータベースの一つである「漢字字体規範データベース」(Hanzi Normative Glyphs; HNG) の CHISE 文字オントロジーとの統合を試みた。ここでは、その概要について述べる。

Categorizing glyph-images based on Multiple Granularity Hanzi Structure Model

— An experimental integration of HNG and CHISE —

MORIOKA, Tomohiko (Institute for Research in Humanities, Kyoto University)

This paper describes about an experimental integration of “Hanzi Normative Glyphs” (HNG) and the CHISE character ontology. The CHISE character ontology uses the Multiple Granularity Hanzi Structure Model to support various glyphs and multiple unification granularity of Chinese characters. This model works fine for modern glyphs of Chinese characters, however it is not clear that the model is sufficient for premodern Chinese characters. In addition, to design reasonable unification rules for each unification granularity, we need various glyph examples of Chinese characters. In these senses, the CHISE character ontology should integrate glyph database and/or glyph corpus. Therefore, we try to integrate HNG and the CHISE character ontology.

1 はじめに

漢字字体の変遷やその規範意識の移り変わりを考える場合、漢字字体規範データベース (Hanzi Normative Glyphs; HNG) や拓本文字データベース [1] といった、漢字字形の用例を収録したグリフデータベースやグリフコーパスは大変有用な道具だといえる。特に、HNG は漢字の包摂規準を設計する上で非常に有用な情報が含まれているといえるが、その背景となる漢字字体の判定規準は十分に機械可読化されているとはいえない。拓本文字データベースでは異体字をある程度統合した上でそれらを代表する UCS 統合漢字で管理しており、どの字体に属するかの情報はない。一方、HNG では主に「大字典」の基準に立脚し、[8] 石塚晴通氏の経験に基づいた字体の整理が行われており、ソース毎の字体の区別の情報が存在するケースもあるが、ソースを跨いだ字体のグルーピングはなされていない。また、拡張 B 以降の統合漢字とのマッピング情報を欠いており、今日的にはやや問題があるといえる。また、「CHISE IDS 漢字検索」[9] のような漢字の部品を用いた検索ができない。そこで、HNG に収録された漢字字形を CHISE で採用している多粒度漢字構造モデル [7] に基づいて整理し、CHISE 文字オ

ントロジーに収録することを試みた。

2　HNG とは

漢字字体規範データベース (Hanzi Normative Glyphs; HNG) [8] 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」は紙カードで整理されていたが、15 年程前から電子化が開始され、2004 年度から豊島正之氏の管理のもとで Web 上での検索サービスの公開が始まった。

前述のように、HNG は「石塚漢字字体資料」を基に構築されたが、後に、典籍の原本や影印本の撮影画像から直接用例を収集したものが追加された。[13] これは、グリフコーパスとしての性格を持つデータといえ、紙カードをベースにした元々の HNG とは性格の異なるものといえる。[1]

3　HNG のデータ構造

2 節で述べたように、HNG には、現在、「石塚漢字字体資料」の紙カードをベースにしたものと、全文画像中の文字を画像マークアップしたグリフコーパスの２つからなっている。本研究では、まず、前者のデータを基に作業を行うことにした。

このデータには、妙法蓮華經卷五（今西本)、妙法蓮華經卷三（守屋本)、開成石經孝經、といったソース毎に文字を切り出した紙カードを電子化したもので構成されており、各ソース毎に、各文字に対応する「石塚漢字字体資料」の紙カードの写真と、それを切り出した各字形の写真が存在する。紙カードは 10 進 4 桁の番号が振られており、それに対応する各字形は異体字が存在しない場合はソースを示す接頭辞に紙カードの番号を付けたものを ID とし、異体字が存在する場合にはそれにさらに a, b, ... といった接尾辞を付けたものを ID とすることで両者の関係が紐付けられている。また、ソースを跨いだ各 ID 間の関係は Excel の表で管理されている。

4　字形整理上の問題

HNG は、版本だけでなく、手書きの写本や拓本も収録しているが、[8] で指摘されているように手書き文字では書き手によって同一の字体であっても個々の字形が著しく異なる場合があり、それらを機械的に別字体とすると意味もなく異体字が爆発してしまい都合が悪い。また、書き間違いの問題もある。拓本の場合、拓本の取り方によって点や線が欠けてしまったり（図 1, 2）余計なゴミが写ってしまう場合があるが、こうしたものも機械的に別字体とするのは問題であるといえる。HNG ではこうした問題に関して、石塚晴通氏らの経験や研究の蓄積に基づいた判断が行われている。が、そうした判断規準自体は必ずしも明文化されておらず、無知識的かつ機械的に判断することは難しい。そういう意味では、HNG のデータから推測される判断規準を勘案して判断する必要があるといえる。

図 1: 拓本の例（開成石經孝經 0257「世（丗)」)

図 2: 拓本の例（開成石經孝經 0011「位」)

[1] なお、http://www.joao-roiz.jp/HNG/ で公開されていた HNG の Web サービスは、2015 年春頃から 11 月現在に至るまで、長期にわたって利用できない状態が続いている。このため、著者は HNG の最新版の本来のありようがどうであったかを正確に記すことができず、HNG について書かれた論文中の記述や著者のあやふやな記憶でしか記すことができず残念である。

5 CHISE 文字オントロジー

CHISE 文字オントロジー [6] は文字処理のために著者らが開発している軽量オントロジーである。CHISE 文字オントロジーは Unicode に収録された文字の情報の他に、漢字に関しては Unicode の包摂規準以外に超抽象文字や字体・字形といった複数の包摂粒度による漢字のグリフに関わる情報を持っている。各漢字には、部首・画数や異体字・類字関係等の情報、IDS 形式 [2] に基づく漢字構造情報、各種文字符号でのコードポイントの情報、各情報の出典等のメタ情報を収録しており、現在のデータ総数は約 24 万オブジェクト（抽象文字、超抽象文字、字体、字形等の各粒度のオブジェクトののべ数)、89 万トリプルである。

CHISE 文字オントロジーは、現在の所、文字に関わる情報だけを収録しており、文字以外のリソースは単なる識別子や外部へのリンクになっている。

6 多粒度漢字構造モデル

多くの漢字は偏と旁などの部品の組み合わせによって構成されている。こうした漢字の部品の組合せ構造に関する情報のことを「漢字構造情報」と呼ぶことにする。漢字構造情報の機械可読な表現法として幾つかの形式が提案され利用されてきたが、[12] Ideographic Description Sequence (IDS) 形式が ISO/IEC 10646 [2] の一部として標準化されている。

漢字構造情報は部品の組合せ方を示すオペレーターと部品からなる構文木で表現できる。IDS はオペレーターとして IDC (Ideographic Description Characters), 部品として UCS の統合漢字および部品用文字を用いたものであるが、部品としてそれ以外のものを用いることも原理的には可能である。

ここで、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 3)。[10]

多粒度漢字構造モデルにおいて、どのような包摂粒度階層を用いるかは随意であるといえるが、現在、CHISE 文字オントロジーでは、主な階層として、超抽象文字（字種）–抽象文字–抽象字体–抽象字形–字形 という 4 階層の粒度を用いている。また、補助的な階層として、抽象文字粒度と抽象字体粒度の間に統合字体粒度、抽象字体粒度と抽象字形粒度の間に詳細字体粒度を置くことを許している。

本稿では、包摂粒度付き文字情報を、超抽象文字は「〈*字*〉」、抽象文字は「〈字〉」、統合字体は「{ 字 }」、抽象字体は「字」、抽象字形は「《字》」、字形は「『字』」のように表記することにする。

7 CHISE での表現

HNG の情報を CHISE 文字オントロジーに取り込むには幾つかの方法が考えられるが、ここでは HNG の各字形を CHISE における字形オブジェクトとして表現し、それを CHISE 文字オントロジー中の既存の抽象字形オブジェクトのどれかに張り付けることにする。

もし、既存の抽象字形オブジェクトのいずれにおいても包摂することができなかった場合、包摂可能な抽象字体オブジェクトの直下、もしくは、新たに抽象字形オブジェクトを設けてその下に張り付けることにする。同様に、もし、既存の抽象字体オブジェクトのいずれにおいても包摂することができなかった場合、包摂可能な統合字体オブジェクトの直下、もしくは、新たに抽象字体オブジェクト（と抽象字形オブジェクト）を設けてその下に張り付けることにする。以下、同様に、抽象文字、超抽象文字と包摂粒度を上げて行き、どの包摂粒度でも包摂できなかった場合は孤立用例とする。

こうすれば、CHISE 文字オントロジー中のいずれかの場所に HNG の字形オブジェクトを位置付けることができる。また、もし、既に存在する抽象字形オブジェクトで包摂可能な場合、漢字構造情報 (IDS) を新たに記述する必

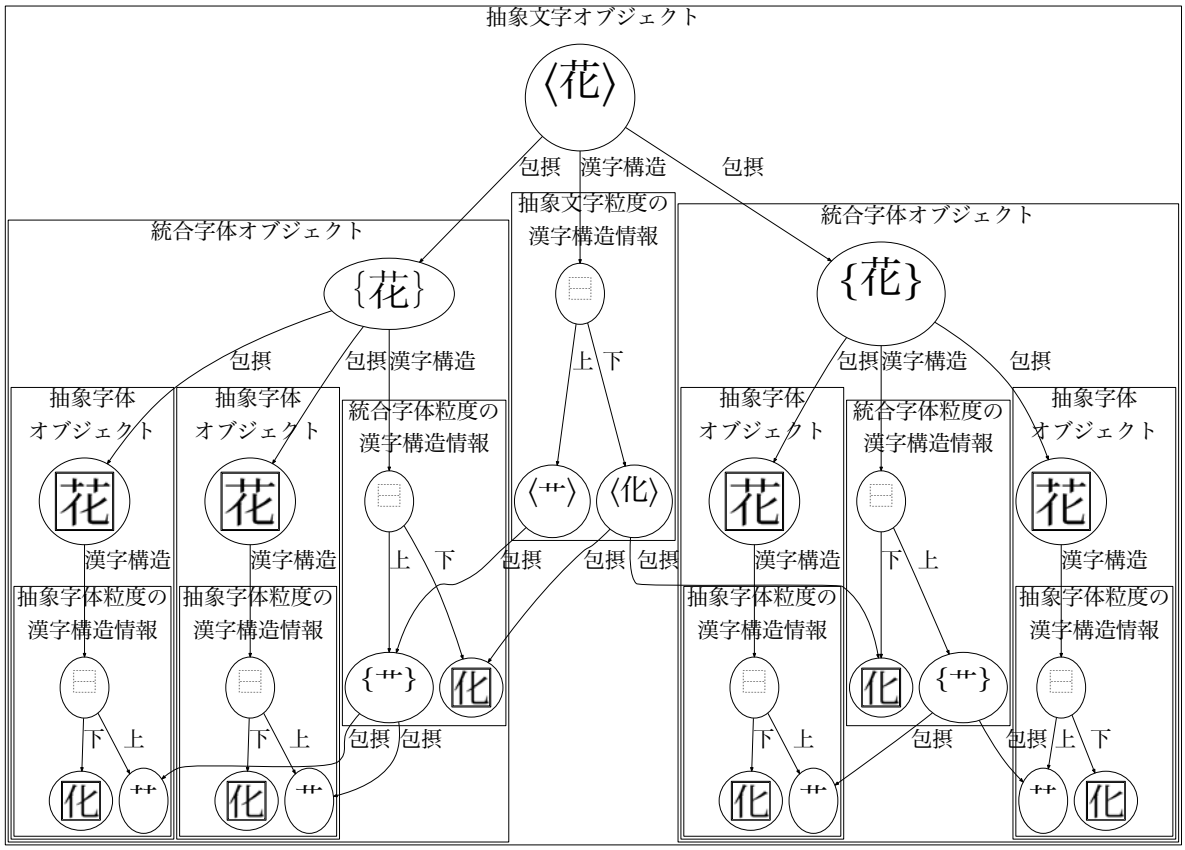


図 3: 多粒度漢字構造モデルの概念図（花）

要がない。

HNG の各字形はそのソース毎に字形粒度の ID 素性と字形の ID で管理することにする。

HNG では各ソースに対し、3 文字のラテン文字からなるソース ID を付けているので、CHISE では字形粒度を示す接頭辞 `===` と HNG を示す `hng-` の後に小文字 3 文字のソース ID を付けて`===hng-`*abc* のように表現することにする。

例えば、開成石經孝經の場合、ソース ID は ‘kak’ であるので、CHISE における ID 素性は `===hng-kak` となる。

一方、字形の ID は、カード番号を 10 倍し、接尾辞がないものは 0, 接尾辞が a のものは 1, 接尾辞が b のものは 2, 以下、接尾辞に対応した番号を足した番号を素性値として用いることにする。

8 包摂規準の問題

現在、CHISE project では、字体・字形粒度の包摂範囲を規定するためのガイドラインとして、「CHISE 文字オントロジーのための漢字字体・字形粒度の情報記述に関するガイドライン (CHISE Guidelines for Glyph Granularity of Chinese characters; CHISE-GGG) Ver.0.9」[11] を策定し、これに則る形に CHISE 文字オントロジーを修訂する作業を行っている。HNG 字形オブジェクトの CHISE 文字オントロジーの取込作業でもこのガイドラインに則って、統合字体、抽象字体、詳細字体、抽象字形の包摂範囲を判定することにする。

また、抽象文字の包摂範囲は、原則として、UCS の統合漢字の符号化作業で用いられている IRG Working Document Series (IWDS) [5] 1: List of UCV (Unifiable Component Variations) of Ideograhs を用いることにする。

但し、IVD (Ideographic Variation Database) [4] に登録されているグリフがその IVS (Ideographic Variation Sequence) [2] の基底文字である統合漢字の IWDS-1 から導出可能な包摂範囲の外にある場合 (つまり、IVS で表現される異体グリフがその親字である統合漢字で包摂できない場合)、その統合漢字の包摂範囲を拡張し (包摂規準を追加し)、IVS で表現されるグリフは全て包摂できるものと看做すことにする。この場合、IWDS-1 の包摂範囲に基づく UCS 抽象文字粒度を示す ID 素性 `=>ucs@iwds-1` を用い、IWDS-1 に基づく抽象文字オブジェクトを表現し、拡張された UCS の抽象文字オブジェクトと IWDS-1 に基づく抽象文字オブジェクトの間の包摂関係を記述することにする。

同様に、IWDS-1 では明示されていないが、抽象文字として包摂した方が良いと思われるケースに関しては、UCS の包摂範囲を拡張することにする。この場合も、`=>ucs@iwds-1` 素性を用いて IWDS-1 に基づく抽象文字オブジェクトを表現し、元々の UCS 統合漢字の抽象文字であったものをこれで置き換えて表現し、拡張された包摂範囲との包摂関係を記述することにする。しかしながら、具体的にどういう場合において包摂規準を拡張して同字とし、どういう場合においては別の抽象文字とするべきか判断に悩むケースが少なくない。[2] ここでは、作業中に見つかった幾つかの例を挙げる。

8.1 包摂できそうなもの

解釈次第では既存の IWDS-1 で包摂可能であると思われるが、当面、UCS の包摂範囲の拡張として扱う。

「少」と「⺌」  (宮廷今西:0043「少」)

8.2 包摂した方が良さそうなもの

厳密には既存の IWDS-1 で包摂できないといえるが、HNG では同じ字体として扱われていると考えられ、UCS 統合漢字の包摂実態から類推して包摂しても問題が少なさそうだと考えられるため、UCS の包摂範囲を拡張する。

8.2.1 書写上の微小なデザイン差

「禺」と「禺」  (宮廷今西:0066「愚」)

「厶」と「𠂤」  (宮廷今西:0395「或（或）」)

8.2.2 筆運び上の省略

「昜」と「𣏌」  (宮廷今西:0321「塲」)

「聶」と「聶」  (宮廷今西:0397「攝」)

「𠬜」と「𠬝」  (宮廷今西:0425「服」)

 (宮廷今西:0032「報」)

「褱」と「𢘑」  (宮廷今西:0374「懐」)

8.2.3 漢字構造の差異に及ぶもの

「□亅米」と「日米亅」  (宮廷今西:0083「断」)

[2] [14] では、明治前期雑誌の漢字の異体字処理において、(1) 既存の基準の拡大解釈で包摂可能なもの (2) 既存の基準に類例が見出せるもの という 2 つのケースの場合に包摂規準の拡張を行うことを原則としているが、実際には線引きが難しい例も少なくないようである。

8.2.4　異体部品が UCS に存在する場合

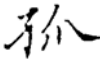
これらのケースの場合、異体部品を単純に包摂すると別字が衝突してしまう可能性があり注意が必要である。

「殳」と「旻」[3]  (宮廷今西:0078「撃」)

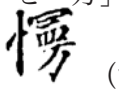
「匕」と「匸」  (宮廷今西:0348「尼」)

「工」と「匸」  (宮廷今西:0632b「差」)

「毎」と「母」  (開成孝經:0013「侮」)

「瓜」と「爪」  (宮廷今西:0039「孤」)

「方」(舅)と「万」(舅)

 (宮廷今西:0061「慢(慢)」)

「臤」と「取」  (宮廷今西:0313「堅」)

「⿰隋土」と「⿰隋工」

 (宮廷今西:0204「堕」)

8.2.5　別部品衝突の可能性があるもの

「舟」と「⺆」  (宮廷今西:0434「槃」)

このケースの場合、「般」と「朌」の両者の包摂とすれば問題がなさそうである。

「攴」/「攵」と「旻」

 (宮廷今西:0405「散」)

「攴」/「攵」と「妥」

 (宮廷今西:0404「數(數)」)

これらのケースの場合、「殳」と部品字形が衝突する可能性があり注意が必要である。

[3]U+22936 に両者の例示字形あり。

8.3　包摂できなさそうなもの

8.3.1　UCS に異体字が存在する場合

「木」と「才」/「扌」

 (宮廷今西:0089「校(挍[4])」)

「釆」と「米」

 (宮廷今西:0377「悉(悉)」

 (宮廷今西:0353「幡(幡)」

「垂」と「𠂤」

 (宮廷今西:0031「垂(𠂤)」

「尼」と「屶」

 (P.2179:0053「尼(屶)」

「壽」と「壽」

 (宮廷今西:0075「擣(擣)」

「念」と「念」

 (宮廷今西:0379「念(念)」

「性」と「忹」[5]

 (宮廷今西:0067「性(忹)」

「惡」と「悪」

 (宮廷今西:0366「惡(悪)」

[4]正字通によれば、「挍」は「校」の忌避字

[5]戸籍統一文字 118240, MJ057495 では異体字としていない。

8.3.2 漢字構造の曖昧性

「侯」と「矦」


(開成孝經:0275「侯 (矦)」)

8.2.3 節の例に似ているが、このケースの場合、別の漢字構造の文字が UCS で符号化されているため、どちらとして解釈するかが文字符号化上の問題になってしまう。[6]

8.3.3 その他

「亡」と「匡」[7]

(兼方紀 2:0330「服」)

「厽」と「亦」

(兼方紀 2:0998b「亦」)

「解」と「觧」

(宮廷今西:0058「懈」)

「雚」と「雀」

(宮廷今西:0444「歡」)

「戎」と「式」

(宮廷今西:0394「戒（戎）」)

「虍」と「𠂢」

(宮廷今西:0072「戲」)

8.4 別字（部品）衝突

「丹」と「舟」[8]

(開成周易:0001「丹」)

「己」と「巳」

(宮廷今西:0633a「己」)

「己」と「已」

(宮廷今西:0633b「己」)

「已」と「巳」

(宮廷今西:0634a「已」)

9 実装

現在、試験的に、妙法蓮華經卷五 (今西本) の 647 字形と妙法蓮華經卷三 (守屋本) の 593 字形を対象に CHISE 文字オントロジーへの取り込み作業を進めている。これまでの所、大半のケースでは既存の抽象字体に包摂可能であるが、IRG [3] において UCS 統合漢字として提案される漢字の重複判定に使われている IWDS-1 [5] の包摂規準では既存の UCS 統合漢字に包摂できない例も若干存在する。[9] こうしたケースの中には手書き文字や拓本といったソースのメディア特性を勘案すれば JIS X0213 や IWDS-1 の包摂規準を拡張するのが自然だと考えられるケースもある一方、判断に困るケースも若干存在した。こうした場合、とりあえず、孤立した字形オブジェクトとして定義し、判断を留保することにした。

このように、HNG の CHISE への収録作業にはそれなりの手間と時間がかかるといえる。そこで、HNG の情報から機械的な変換可能な部分だけを使って文字定義を行い、HNG の字形オブジェクトとして定義することにした。また、HNG における UCS とのマッピング情報（あるいは、大漢和とのマッピング情報）を用い、既存の CHISE 文字オントロジーの UCS の抽象文字オブジェクトから HNG の字形オブジェクトに対して関係素性 `->HNG` を張ることにした。これにより、未整理の字形もとりあえず CHISE-wiki で表示させることができ、CHISE IDS 漢字検索の恩恵もある程度利用可能になるといえる。

[6]「隹」も同様に分離して書かれるケースが多々あるが、分離した場合の漢字構造に該当する別字がないので曖昧性が生じにくいようである。

[7]戸籍統一文字 002080, MJ056865

[8]汚れか?

[9]IRG に提案中のものもあった。

10 おわりに

HNG は石塚晴通氏らの経験や研究の蓄積を反映した貴重な漢字グリフコーパスのひとつであるといえるが、この背景となる漢字自体に関するさまざまな知識そのものは多分に暗黙知を含んでいるといえる。そういう意味では、ここで行っている作業は HNG に含まれている暗黙知を CHISE に翻訳することで結果的に機械可読化する作業という風にとらえることができるかもしれない。

最後に、こうした機会を与えて頂いた、豊島正之先生、高田智和先生、そして、石塚晴通先生に感謝する。なお、本論文における誤りや誤解は全て私の責任であることはいうまでもない。

参考文献

[1] 拓本文字データベース. http://coe21.zinbun.kyoto-u.ac.jp/djvuchar.

[2] International Organization for Standardization (ISO). *Information technology — Universal Coded Character Set (UCS)*, 2014 年 9 月. ISO/IEC 10646:2014.

[3] ISO/IEC JTC1/SC2/WG2/IRG (Ideographic Rapporteur Group). http://www.cs.cuhk.edu.hk/~irg/.

[4] Ideographic Variation Database. http://unicode.org/ivd/.

[5] IRG Working Document Series. http://appsrv.cse.cuhk.edu.hk/~irg/irgwds.html.

[6] Tomohiko Morioka. CHISE: Character processing based on character ontology. In *Large-scale Knowledge Resources (LKR2008)*, No. 4938 in LNAI, pp. 148–162, 2008 年 3 月.

[7] Tomohiko Morioka. Multiple-policy character annotation based on CHISE. *Journal of the Japanese Association for Digital Humanities*, Vol. 1, No. 1, pp. 86–106, 2015 年 11 月.

[8] 石塚晴通, 池田証寿, 岡墻裕剛. 漢字字体規範データベースとその応用. 東洋学へのコンピューター利用 第 17 回研究セミナー, 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. 53–63, 2006 年 3 月.

[9] 守岡知彦. CHISE IDS 漢字検索. http://www.chise.org/ids-find.

[10] 守岡知彦. CHISE に基づくグリフ・オントロジーの試み. じんもんこん 2009 論文集, 情報処理学会シンポジウムシリーズ, 第 2009 巻, pp. 9–14. 情報処理学会, 情報処理学会, 2009 年.

[11] 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ー利用 第 26 回研究セミナー, 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, pp. 153–190, 2015 年 3 月.

[12] 守岡知彦, クリスティアン・ウィッテルン. 文字データベースに基づく文字オブジェクト技術の構築. 情報処理振興事業協会平成 13 年度 成果報告集. 情報処理振興事業協会, 2002 年. http://www.ipa.go.jp/NBP/13nendo/reports/explorat/charadb/charadb.pdf.

[13] 高田智和. 漢字字体と典籍の性格との関係—「漢字字体規範データベース」が主張するもの—. 情処研報, Vol. 2013-CH-97, No. 12, pp. 1–4, 2013 年 1 月.

[14] 須永哲矢, 堤智昭, 高田智和. 明治前期雑誌の異体漢字と文字コード —『明六雑誌』を事例として—. じんもんこん 2011 論文集, 情報処理学会シンポジウムシリーズ, 第 2011 巻, pp. 381–388. 情報処理学会, 情報処理学会, 2011 年.