データベースの再生と保存についての試論 —HNG を例に—

守岡 知彦 （京都大学 人文科学研究所）

人文情報系データベースを長期間維持することの困難さが認識されるようになってきたが、実際にサービスが停止してしまったデータベースを復元しそのデータを将来にわたって維持することは必ずしも容易ではないといえる。ここでは、漢字字体規範史データベース (Hanzi Normative Glyphs; HNG) の分散型版管理の利用したデータセット化、研究者の所属機関や営利企業のプラットフォームに依存しない Git サービスの提供、データセット保存会といった漢字字体規範史データセットに関する取組みについて概説するとともに、人文情報系データベースの長期保存にかかわる問題についても併せて議論する。

An Essay on Database Recovery and Preservation — Based on the case of HNG —

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

Although the difficulty of keeping databases on digital humanities for a long time has come to be recognized, it seems that the feasible method for restoring the database whose service has actually stopped and maintaining the data in the future is not yet well established. In this paper, we will outline the efforts on the HNG dataset, such as publication of dataset using distributed version control system (Git), provision of Git hosting service independent of URL of researcher's institution or platform provided by commercial companies, organization of dataset preservation association, and also discuss issues related to long-term preservation of databases.

1 はじめに

人文情報系データベースを長期間維持することの困難さが認識されるようになってきたが、実際にサービスが停止してしまったデータベースを復元しそのデータを将来にわたって維持することは必ずしも容易ではないといえる。ここでは、漢字字体規範史データベースを例にこの問題について議論する。

2 漢字字体規範史データベース

漢字字体規範史データベース (Hanzi Normative Glyphs; 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」は紙カードで整理されていたが、15 年程前から電子化が開始され、2005 年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。

HNG は収録する資料に含まれる字形用例を字体に分類し、その代表字形を字種（抽象文字）によって管理し、字種粒度によるソースをまたいだ串刺し検索・一覧表示を実現することにより、文字（字種単位）での字体の変遷を把握可能にしている。また、標準・規範的な度合が強い公的な写本・版本（および、石刻文字）を中心に私的な写本・版本等も比較のために収録することで、楷書の字体規範の時代的・地域的変遷と異化の全体像を把握することを可能にしている。特に、敦煌本を始めとする唐代以前の中国古写本と、奈良・平安時代の日本古写本を通して、初唐の標準字体が日本の標準字体として移入・定着する様

相を精緻に記述する基盤を提供している。また、開成石經の字体が宋版を通じて受容されることによって初唐の標準字体とは異なる字体が新たな規範字体として定着する様子を描写している。

こうした HNG の特徴は石刻拓本文字データベースと対照的であると評することができる。即ち、石刻拓本文字データベースが未整理の全用例を提示する漢字字形コーパスであるのに対し、HNG は資料選定の段階において石塚晴通氏の学識に基づいて典型的な標本が選択されており、また、字形が字体に分類され、資料毎に各字体を代表する（なるべく綺麗で判りやすい）例示字形が選択されており、石刻拓本文字データベースの検索結果と対照的な必要最低限の少ない数の字形用例を出力できるようにデザインされていた。

前述のように、HNG は、本来、漢字字体史研究における幾つかの仮説を実証するという特定の研究目的のために作成されたデータベースであったが、比較的少ない標本点と簡単な UI によって漢字の楷書字体の変遷を把握できるように設計された HNG は漢字字体史を専門としないユーザーにとっても有用なツールとなり、次第に基盤的な漢字データベースの一つとして広く使われるようになっていった。[*1] しかしながら、HNG の運用体制は必ずしもそうしたインフラとしてのデータベースを長期間維持するものではなかったと考えられる。そうした中、10 年という長期にわたってデータベースサービスが維持されたことは高く評価できるが、データベースの永続化という点では幾つかの問題があったと考えられる。

3 HNG の別実装の開発

2014 年 9 月に、科研費基盤研究 (B)「字体記述のデジタル化に基づく文字規範史の定位」の一貫として、CHISE と HNG の連係に関するプロジェクトが始まった。これは、CHISE の技術を用いて HNG の例示字形の字体記述を行うことを目指したもので、豊島正之氏のアドバイスに基づき、長安宮廷写経を対象に作業を始めることとなり、2015 年 2 月に高田智和氏から今西本妙法蓮華經卷五と守屋本妙法蓮華經卷三のデータを頂いてこれらの CHISE 文字オントロジーへの統合に関する検討を始めた。ただ、年度末ということもあり、具体的な作業を始められたのは 2015 年度に入ってからのことである。

HNG が公開から 10 周年を迎えた 2015 年の 4 月頃[*2]、HNG はサービスを停止した。2015 年 11 月には HNG 公開 10 周年を記念するイベントが企画されており、その発表者は HNG が利用できない状況で HNG について語らざるを得ないという状況に陥った。著者にとっても、今西本妙法蓮華經卷五の CHISE 文字オントロジーへの統合のための作業を始めようとした矢先の出来事であり、その停止の長期化は困った事態であった。

そこで、著者は高田智和氏の手元にあった HNG の古いバックアップデータを元に HNG 代替サービスとして CHISE-wiki 上での HNG 字体資料の公開を始めることとなった。これは今西本妙法蓮華經卷五と守屋本妙法蓮華經卷三（後に、開成石經論語）を対象に多粒度漢字構造モデルに基づく精緻な字体記述を行う一方、残りの資料に関しては HNG の見出し字の情報を用いて対応する UCS 抽象文字オブジェクトに HNG の例示字形一覧を張り付けるという簡易的な対応を行うことにより、短期間に HNG 全資料の公開を実現しようとする試みであった。[6] この方法により、2015 年 10 月には当時手元にあった HNG 48 資料の公開を完了し、また、11 月には「CHISE-IDS HNG 漢字検索」を公開した。[8]

また、HNG 公開 10 周年イベントの場で、ユーザーから「石塚漢字字体資料」の紙カード画像を公開して欲しいという要望があり、石塚晴通氏他関係者が一般公開することで合意したため、2016 年 4 月に CHISE-wiki に石塚漢字字体資料の紙カード画像の表示機能を追加し、2016 年 6 月には IIIF Image API を利用した紙カード画像と開成石經拓本画像の配信、及び、これを利用した京大人文研所蔵の開成石經画像と HNG の開成石經データの比較表示機能も追加した。[7] これにより、48 資料に限られるものの HNG の情報が再びインターネット上から自由に利用できるようになり、また、「CHISE-IDS HNG 漢字検索」を用いた部品単位での検索や比較、紙カード画像に戻っての再検討、HNG の開成石經の拓本

[*1] 高田智和氏の言葉を借りれば「研究用データベースのインフラ化」が起こったといえる。

[*2] 正確な日時は良く判らない。

よりも良い拓とされる京大人文研所蔵の開成石經画像との連携機能といった従来にはなかった高度な機能が追加され、CHISE やそれと相互リンクしている GlyphWiki や UniHan データベースといった新たな導線ももたらされた。その一方で、旧来の UI が使いたいユーザーの要望を十分にかなえたものとはなっていなかった。

4 資料の発掘とデータセット化

コンピューターシステムには、これまで、数年に一度位に、大きなアーキテクチャー変化の波がやって来ており、10 年前のネットワーク環境と今日のそれは異なっているし、また、10 年後に今日の Web 環境やモバイル環境がそのままの形で続くかどうか疑わしく、デバイスの変化、ネットワークインフラの変化、ソフトウェア技術の変化、社会的な変化等を背景に今後もコンピュータ環境は変化し続けると考えられる。

長期にわたってデータベースや情報サービスを維持するためにはこうしたアーキテクチャー変化の波を乗り越える必要があり、それを乗り越えられなかったシステムはサービスを維持できなくなる。あるいは、かろうじてサービスを維持しても、古臭いインターフェースしか提供できずに使いにくくなってしまうこともあるだろう。コンピューター・アーキテクチャーは変化して行くので、設計時に普及していた製品や技術、運用体制や使われ方といったもろもろの前提が数年後には崩れてしまうということも少なくない訳である。

こうした問題に対処する方法の一つに、データベースから Web 上での検索サービスや UI といったプログラム的な部分を除いたデータそのものをデータセットとして公開するという方法が考えられ、東寺百合文書や日本古典籍データセットのオープンデータとしての公開が注目を集めた。HNG もその長期安定的なデータ公開と計算機環境の変化に合わせた今後の発展を鑑みれば、従来の HNG で公開していた漢字字形の切り抜き画像とメタデータに加え、石塚漢字字体資料の紙カード画像と聞き取り調査やデータ発掘等で得られた知見に基づくメタデータやオントロジー、文書等をデータセットとして公開することが望ましいと考えるに至った。「漢字字体規範史データセット」[9] はこの立場に基づき、HNG の主要部分を Git リポジトリ化し、オープンデータとして公開することを目指したものである。

4.1 データの発掘と整理

3 節で述べた HNG の別実装の開発において、先ず行ったのは当時手元にあった HNG 48 資料（おそらく、2007 年 3 月頃のデータだと思われる）のデータの抽出と解析、及び、整理されたデータの Git リポジトリ化であった。

後に、北海道大学文学研究科池田証壽研究室に残されていた 2010 年 3 月頃のバックアップデータを頂き、これに 64 資料版のデータが含まれていることが判明した。ただ、このバックアップデータには複数の時期に複数の作業者が作成したと思しき、重複したり異同があったりデータ形式の変更や画像の撮り直し等も存在しており、どれが最新かつ適切なデータか把握しづらいものであった。

HNG では資料に対して番号と ASCII 3 文字による略称（他にも、文献名と人間向けの略称が存在する）が付与されているが、この番号は基本的にメタデータを記した Excel ファイルにおいて自動生成されたものだったようで時期により変化していた。また、これとは別に各資料のフォルダー名の接頭辞として使われている番号もありこれはおそらくある時期の Excel ファイルに対応するものと思われるが、結果的に、資料通番とフォルダー番号という 2 系統の番号に分化しているらしいということが判明した。一方、ASCII 3 文字による略称は比較的安定しており ID と見なせそうだが、こちらも若干の異同の存在が判明し、各版の比較検討が必要だということが判った。

こうしたことから、64 資料版 HNG のデータを確定することはすぐにはできず、当面、48 資料版の Git リポジトリをベースに作業を行っている。

4.2 仕様や用語の調査

一般に一度止まってしまったサービスを復元するのは難しいといえるが、この要因の一つは元々のサービスの現物が見れないため、その挙動を厳密に知ることができないことにあるといえる。

HNG の場合、そのバックアップデータからサービスの元となるデータを入手することができたが、版管理されていない複数の時期のさまざまなバリエー

ションが混在したものであったために、どれが正しいデータか判らないという問題が存在した。

また、データモデルや用語の定義が良く判らず、また、仕様が不明なためにどのようなデータがどのように検索されどのように表示されるか、言い替えれば、データのセマンティクスが十分に形式化・文書化されていないために、論文等で文書化された情報や関係者からの聞き取り調査等に基づき推測する必要があった。しかしながら、論文等ではデータやシステムの詳細な仕様は書かれておらず、また、発表時期によってそれぞれ異なった版の HNG について述べているという問題もある。また、関係者といえども全てに関っている訳ではない上、記憶違いや忘却も起こるため、これまた完全ではない。とはいえ、論文等で発表されていないことがらの意図について知る上で極めて重要な情報といえる。

そのため、こうした聞き取り調査の結果を（人間向けに）文書化するとともに、機械可読な形でも表現することに取り組んでいる。ここでの対象となる事項にはデータ形式や検索システムの仕様といった比較的計算機システムよりのものの他、各資料に関する情報やその選定理由等、あるいは、「書体」や「字体」、「標準」、「公的」、「私的」といった概念も重要である。こうした概念は使用者によって揺れがあったり石塚漢字字体資料や HNG において独特の使われ方やニュアンスをおびている場合があり、聞き取り調査や用例、他の概念との関係等によって（石塚晴通が元々どういう意味づけをしていた・どういう意図で使っていたかや石塚漢字字体資料・HNG でどういう運用がなされてきたか等を中心に）明確化することを目指している。この一部は、[4] や [5] として既に公開されている他、今後もその調査や記述を進め、Git リポジトリ、Web サイト、論文・書籍等の文書等で公開して行く予定である。

一方、HNG の内容やモデルを理解する上で、石塚漢字字体資料や HNG の作成過程を知ることは重要であると思われる。特に、データに疑義がある部分が見つかった時、元資料に当たることも重要であるが、その解釈や分類等の HNG 固有の部分が問題となった時や、対応する元資料へのアクセスに問題があるような場合には石塚漢字字体資料の紙カードを保存することも重要であると考えられる。このため、北海道大学文学研究科池田証壽研究室の協力を得て、その資料調査を行った。

4.3 従来型検索 UI の再現

3 節で述べたように、現在、CHISE-wiki 上での HNG 字形表示機能と紙カード画像表示機能、および、「CHISE-IDS HNG 漢字検索」を用いた部品による HNG 漢字検索機能により Web 上で HNG のデータを検索・閲覧することができるが、旧来の UI が使いたいというユーザーの声も寄せられており、オリジナルの HNG に近い UI を再現することを計画している。

オリジナルの HNG では、そのデータ公開においてクローラー等がその内容を全部取って行くことを懸念して、わざと検索結果や画像の URL が一意にならないような工夫を行っていたようであるが、現代においてはむしろ検索結果に対してパーマリンクを設けることが必須といえ、長期保存やサービス停止後の復元という観点でもこのようなアクセスに障害を設けるような工夫を行わないことが重要であるといえる。また、スマートフォン等での利用を考慮することも重要であろう。このため、オリジナルの HNG に近い UI の再現においては、パーマリンクとレスポンシブデザインを実現したいと考えている。

ところで、オリジナルの HNG が停止して既に 3 年余りが経っており、実の所、オリジナルの HNG の挙動に関する記憶は徐々に風化しつつあるといえ、その仕様を確定することは実の所容易ではない。また、上述のように、オリジナルの HNG を完全再現することが良いとは必ずしもいえず、そのエッセンスを残しつつ妥当で実現の容易な仕様を確定することが重要であるといえる。

この際、重要なヒントの一つは HNG のスクリーンショットである。論文や学会発表等でのスライド、記事等で引用される形で断片的に残っているが、HNG 自体何度かバージョンアップを重ねており検索結果も変化しているため、いつの時期のものかが網羅的に判ることが望ましい訳であるが、こうした引用では着目する部分だけをトリミングしていたりいつの時期・どの版のものを参照しているかが判りづらかったりするなどスクリーンショットの網羅的保存という理想に比べて現状は非常に厳しいものがあるといわざるを得ない。逆にいえば、現在動いている Web

図 1　オリジナル HNG での表示例

サービスの将来に向けての保存という観点では、スクリーンショットの網羅的保存は重要な検討項目ということができるだろう。

オリジナルの HNG の検索結果の表示画面では、上から順に中国写本、中国版本（石経を含む）、日本資料、その他という順に多段表示されていた（図 1）が、HNG データセットの各資料に付けられた「初唐写本」や「日本版本」といった資料の分類に関する項目（[5] では「区分」としている）はこの機能を実現するための情報として用いられていたと考えられる。しかし、この項目の値は「大和寧写本」や「北宋版」などのように、中国写本、中国版本、日本資料、その他という分類とは見掛け上一致していない。

一方、CHISE-wiki での HNG 字形の表示ではこれらを全てまとめて表示していたが、おそらくはこのことがユーザーにとって一番の問題であったと考えられる。CHISE では見出し文字と HNG 字形の対応関係を HNG の見出し文字に対応する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` を張ることで表現していた。[6][8] CHISE-wiki ではこの関係素性 `->HNG` をその標準機能によって単純に順番に表示していたため、HNG 字形が順番にまとめて表示されていた訳である（図 2）。

オリジナルの HNG 風の検索結果を実現する場合、全てを関係素性 `->HNG` にまとめるよりも各資料の区分に基づいたドメインに分けた方が良いといえる。仮に全てをまとめて表示するにしてもドメインに分けた値を混ぜる方が 1 つにまとめたものを分類するよりも簡単だと考えられる。そこで、中国写本、中国版本（石経を含む）、日本写本、日本版本、韓国資料、その他にドメインを分けることにし、それぞれ `->HNG@CN/manuscript`, `->HNG@CN/printed`, `->HNG@JP/manuscript`, `->HNG@JP/printed`, `->HNG@KR`, `->HNG@MISC` というドメイン付き関係素

図 2　従来の CHISE-wiki での表示例

図 3　現在の CHISE-wiki での表示例

性で表現することにした（図 3）。表 1 に HNG データセットにおける資料の区分との対応を示す。なお、従来、日本資料としてまとめられていたものを日本写本と日本版本の 2 つに分けたのは HNG が設計された頃と比べその後の増補により日本資料の数が増えたことと開成石経規範の影響に写本と版本で差があるケースを考慮したためである。

ところで、オリジナルの HNG 風の UI の仕様を定義することはデータセットの項目を明確化する上でも意義のあることだと考えられる。即ち、ある項目は何の目的も無しに漫然と設けられているのではなく、検索や表示といった検索サービス上でのなんらかの挙動を実現するために設けられている訳であ

区分	ドメイン	種別
南北朝写本	CN/manuscript	中国写本
隋写本	CN/manuscript	中国写本
初唐写本	CN/manuscript	中国写本
則天写本	CN/manuscript	中国写本
盛唐写本	CN/manuscript	中国写本
高昌写本	MISC	その他
吐蕃写本	MISC	その他
大和寧写本	MISC	その他
開成石経	CN/printed	中国版本
北宋版	CN/printed	中国版本
南宋版	CN/printed	中国版本
西夏版	MISC	その他
日本写本	JP/manuscript	日本写本
日本版本	JP/printed	日本版本
日本書紀写本	JP/manuscript	日本写本
韓国写本	KR	韓国資料
韓国印刻本	KR	韓国資料

表 1　資料の区分とドメインの対応

り、そうした挙動を明記することはデータの持つ意味（の少なくとも一部）を示すことにつながるといえる。逆に、検索システムのプログラムから見た場合、どういうデータを使ってどう処理するかを示すことはその仕様の一部であろう。

もし、関数型言語でデータフローを示したり論理型言語で推論によって検索を実現するというような宣言的プログラミングでシステムを記述できるなら、そのシステムの記述はデータと見なすことができるであろう。もちろん、現実の関数型言語や論理型・制約型言語で書かれたプログラムは現実の実装を反映し、システム依存の記述を含んだものとなることもしばしばだといえるが、適切なドメイン固有言語を実現することができれば、プログラム、即ち、データの意味論における動的な部分を（なるべくコンパクトな）宣言的記述によって表現することができるかも知れない。

HNG の場合、前述したような各資料の区分と検索結果の表示画面での多段表示の関係は表 1 に示すような各資料の区分と段の写像として表現できるし、同一資料に複数字体が存在する場合の各例示字形の配置法は「字体数」の項目を使って定義できる。もちろん、実際のシステムにはより低レベルの部分が含まれるが、HNG としての意図を形式的に示す上でそうした詳細な実装は必ずしも必要であるとはいえず、HNG 的に意味のある部分は相当程度宣言的に記述可能であると考えられる。

5　データセットの保存と利活用

データベースの永続化を阻む要因は幾つか考えられるが、その理由のひとつは、これらの多くがプロジェクト型競争的資金によって開発され、プロジェクト終了後に予算的・人員的問題から十分な運用体制を取れないからではないかと考えられる。このため、プロジェクト終了後は少数の関係者の努力と熱意（とお金）に依存してしまい、機械の故障やセキュリティー対策、ソフトウェアのバージョンアップ等の問題が生じた時に（担当者が燃えつきてしまい）対策することができないままやむなくサービスを終了してしまったりする。また、運用組織の改組や研究者の移籍等によって受け皿を失ってしまったシステムもあったかも知れない。

こうしたことを鑑みれば、お金がなくても持続できるような、言い替えれば、競争的資金に依存しなくてもすむような仕組みを実現することが重要であるといえる。そのためにはメンテナンスコストを下げるための工夫が必要であり、そのためには、定期的なリファクタリングやその時々のリーズナブルな計算機環境に合わせた改修が必要になるといえる。つまり、逆説的ではあるが競争的資金に依存しない体制を採るためには、多くの場合、競争的資金を使ってその維持管理体制を改修する必要があると思われる。HNG の場合も同様であり、リファクタリング作業の他、物理的な資料の調査・整理、聞き取り調査、ライセンス的に怪しい部分の作り直し等にはそれなりのコストがかかるため、CHISE との統合や IIIF による画像公開の流れを背景にした HNG のリファクタリングとその Git リポジトリ化、および、HNG データセットの長期に渡る安定的な公開体制の確立を目的とした科研費の申請をした所採択され[*3]、実際にその実現に向けて動くことができた訳である。

[*3] 基盤研究 (C)「字体記述の精密化手法の確立による歴史的漢字字体情報アーカイブズ構築」(18K00611)

しかしながら、前述のように、中長期的にはあまりお金がなくても維持できるような体制をとることが重要であり、最悪の場合、利用者の小額の寄付によって維持できるぐらいの体制にすることが望ましいと考えられる。

一方、3 節で述べたように、オリジナルの HNG が停止した約半年後には CHISE で HNG 関連サービスを提供したが気がつかない人が多かった。これは多くのユーザーの元にこの情報が届かなかったからだと思われる。こうしたことを鑑みれば、その正式な配布元となる Web サイトを設けてその存在を周知徹底することが重要であると考えられる。また、この Web サイトの URL は研究者が所属する組織の改組や研究者の移動等によって変化しないようにするべきであり、このため、`hng-data.org` という独自のドメインを確保してその上で Web サービスを行うようにした。また、「漢字字体規範史データセット保存会」を設立し、その設立イベントを通じて Web サイトの広報活動を行った。

現在、Git リポジトリのホスティングサービスとしては GitHub が普及しているが、あえて、独自ドメイン上で GitLab Community Edition を用いて独自のホスティングサービス `gitlab.hng-data.org` を立ち上げたのは、営利企業の提供する占有的なプラットフォームに依存する危険性を回避したかったからである。こうした占有的なプラットフォームは無料で利用できたとしても、ある時大きくポリシーが変わって不都合が生じたり、サービスが停止したり、致命的にその内容が変化する可能性があり、データセットを長期間安定的に提供するという観点では一時配布元としては問題があるといえる。

しかしながら、URL を長期間維持するということ自体の困難さを考えれば、長期的には URL に依存しないアーキテクチャーを考慮することも重要であろう。このため、P2P ベースの分散型ファイルシステムの一つである IPFS (InterPlanetary File System) [1] [2] の利用も検討している。

6 おわりに

データベースの再生のためにはデータ考古学的な手法による『発掘調査』に加え、『データ文献学』や『データ思想史』とでもいうようなデータを解釈する上でのさまざまな観点に基づくデータの『資料批判』や『校訂』といった作業が必要になるといえる。そして、今後、データセットを継承していくためには、今日とは異なるかも知れない計算機環境の上で（自動的に）データを十分に解釈・翻訳可能なセマンティクスを付与する必要があると思われる。これは言い替えれば、データベースを永続化するためにはデータセット本体を保存するだけでなく、データを解釈し処理するプログラムも保存しなければならないという風にとらえることができるかも知れない。そのためには、プログラムをなるべく抽象的かつ宣言的にデータとして扱うこと、即ち、プログラムとして解釈可能なデータという視点が必要になってくると考えられる。

最後に、高田智和氏、石塚晴通氏、豊島正之氏、池田証壽氏、斎木正直氏、劉冠偉氏、北海道大学文学研究科池田証壽研究室の諸氏に感謝する。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] Juan Benet. IPFS - content addressed, versioned, P2P file system (draft 3). *arXiv preprint arXiv:1407.3561*, 2014 年.

[2] Protocol Labs. IPFS is the distributed web. `https://ipfs.io/`.

[3] 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.

[4] 石塚晴通, 高田智和. 漢字字体と文献の性格との関係—「漢字字体規範史データベース（石塚漢字字体資料）」の文献選定. 石塚晴通監修, 高田智和, 馬場基, 横山詔一（編）, 漢字字体史研究 二—字体と漢字情報, pp. 349–359. 勉誠出版, 2016 年 11 月.

[5] 石塚晴通, 高田智和, 守岡知彦. 漢字字体規範史データセット 資料一覧. `http://www.hng-data.org/sources.ja.html`, 2018 年 7 月.

[6] 守岡知彦. 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 —. じんもんこん 2015 論文集, 情報処理学会シンポジウムシリーズ, 第 2015 巻, pp. 1–8. 情報処理学会, 情報処理学会, 2015 年.

[7] 守岡知彦. CHISE-wiki における HNG カード画像利用の試み. 情処研報, Vol. 2016-CH-111, No. 4, pp. 1–8, 2016 年 7 月.

[8] 守岡知彦. CHISE による HNG データ収録の試み. 石塚晴通監修, 高田智和, 馬場基, 横山詔一（編）, 漢字字体史研究 二— 字体と漢字情報, pp. 185–203. 勉誠出版, 2016 年 11 月.

[9] 守岡知彦, 高田智和, 石塚晴通, 他. 漢字字体規範史データセット. https://gitlab.hng-data.org/HNG/hng-data/, 2018 年 9 月.