東方學報 京都第 94 册（2019）：320-284 頁

漢字字體規範史データセット及びその CHISE との統合について

守 岡 知 彦

1 は じ め に

漢字字體規範史データベース（Hanzi Normative Glyphs ; HNG）［14］は時代や地域毎の漢字字體の標準の存在とその變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 年程前から作成を續けてきた字體資料（「石塚漢字字體資料」と呼ぶ）である。「石塚漢字字體資料」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 上での檢索サービスの公開が始まった。

HNG は收錄する資料に含まれる字形用例を字體に分類し，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檢索・一覽表示を實現することにより，文字（字種單位）での字體の變遷を把握可能にしている。また，標準・規範的な度合が强い公的な寫本・版本（および，石刻文字）を中心に私的な寫本・版本等も比較のために收錄することで，楷書の字體規範の時代的・地域的變遷と異化の全體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國古寫本と，奈良・平安時代の日本古寫本を通して，初唐の標準字體が日本の標準字體として移入・定着する樣相を精緻に記述する基盤を提供している。また，開成石經の字體が宋版を通じて受容されることによって初唐の標準字體とは異なる字體が新たな規範字體として定着する樣子を描寫している。

こうした HNG の特徵は石刻拓本文字データベースと對照的であると評することができる。卽ち，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對し，HNG は資料選定の段階において石塚晴通氏の學識に基づいて典型的な標本が選擇されており，また，字形が字體に分類され，資料毎に各字體を代表する（なるべく綺麗で判りやすい）例示字形が選擇されており，石刻拓本文字データベースの檢索結果

と對照的な必要最低限の少ない數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，本來，漢字字體史研究における幾つかの假說を實證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，比較的少ない標本點と簡單な UI によって漢字の楷書字體の變遷を把握できるように設計されたために漢字字體史を專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廣く使われるようになっていった[1)]。しかしながら，HNG の運用體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價できるが，データベースの永續化という點では幾つかの問題があったと考えられる。

2　CHISE に基づく HNG 再生の試み

CHISE（CHaracter Information Service Environment）は著者が中心となって開發している知識處理的アプローチに基づく文字處理環境である。CHISE では，セマンティック Web で使われている RDF［11］と同樣な，有向グラフ型データベース處理系とその上で文字に關する知識を記述するための語彙を使って文字に關するさまざまな知識を記述し，そうして記述された文字に關する機械可讀な知識表現（文字オントロジー）を使って文字を處理する仕組みを提供している。これはいわば UCS 等の符號化文字集合に對するメタシステムに相當するものといえる。CHISE ではこうした文字知識處理のための枠組だけでなく，その上で記述された大規模な文字データ（CHISE 文字オントロジー）も提供している。特に，漢字の部品組合せ方の情報（漢字構造情報）のデータは Web サービス「CHISE IDS 漢字檢索[2)]」でも利用可能であり，UCS の統合漢字をほぼ網羅する檢索サービスとして利用されている。

CHISE 文字オントロジーでは漢字に對して主に文字や文字符號に關わる情報を中心にデータ整備を進めてきたが，漢字を適切にとらえ記述するためには，漢字を讀み書きする人たちの解釋共同體のありようやその規範意識の變遷に着目することが重要といえる。そのためには各時代・地域の字體用例を收集し文字オントロジーに紐付けることが重要であり，こうした觀點から眺めた時，各時代・地域における漢字字體の標準とその變遷を觀察するために設計された HNG は，單にさまざまな漢字字形用例を收錄したグリフ

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）　http : //www.chise.org/ids-find

データベースであるというだけでなく，その資料選擇等で表現されているその基本的なコンセプト自體が漢字の解釋共同體の規範意識を觀察するための實證的なデータを提供しているという意味で極めて重要なものだといえ，CHISE 文字オントロジーを補完するための漢字字體用例データベースとしてふさわしいものだと考えられる。

一方，HNG の側から CHISE を見た場合，「CHISE IDS 漢字檢索」等での部品を用いた漢字の檢索機能や，その檢索結果からリンクされた CHISE-wiki［20］［21］で CHISE のデータを閲覽したり，さらにまた，そこからリンクされた Unihan データベース［7］，グリフウィキ［12］[3]，古典中國語形態素用例データベース［22］，東洋學文獻類目といった他のデータベースにたどっていけること，つまり，文字を核としたポータルサービスとしての機能が有用であるかもしれない。また，CHISE の漢字構造情報のデータを利用して HNG における漢字部品の變遷を見るといったことができたら便利かもしれない。

こうした觀點から，CHISE と HNG の密接な連係・統合には潛在的な利點があると考えられ，2014 年 9 月に，科研費基盤研究（B）「字體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 の連係に關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 の例示字形の字體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，長安宮廷寫經を對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關する檢討を始めた。ただ，年度末ということもあり，具體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[4]，HNG はサービスを停止した。その約半年後の 2015 年 11 月には HNG 公開 10 周年を記念するイベントが控えており，その發表者は HNG が利用できない狀況で HNG について語らざるを得ないという狀況に陷った。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來事であり，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 字體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）を對象に多粒度漢字構造モデルに基づく精緻な字體記述を行う一方，殘りの資料に關しては HNG の見出し字の情報を用いて對應する UCS 抽象文字オブジェクトに HNG の例示字形一覽

3） http : //glyphwiki.org
4） 正確な日時は良く判らない。

を張り付けるという簡易的な對應を行うことにより，短期間に HNG 全資料の公開を實現しようとする試みであった。[24] この方法により，2015 年 10 月には當時手元にあった HNG 48 資料の公開を完了し，また，11 月には「CHISE-IDS HNG 漢字檢索」を公開して [25] 記念イベントを乘り切ることができた。

2.1 データの發掘と整理

CHISE に基づく HNG の別實裝の開發において，先ず行ったのは當時手元にあった 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。

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

・10_妙法蓮華經卷五（今西本）
・01_誠實論卷八（P. 2179）
・17_開成石經孝經

といったソース毎のフォルダーが存在し，この各フォルダーには JPG, BMP, PNG というサブフォルダーが存在する。JPG には各文字（字種）に對應する「石塚漢字字體資料」の紙カードの寫眞が JPEG 形式で格納されている。また，BMP と PNG には紙カードから字體毎に切り出された代表字形の寫眞（漢字グリフデータ）がそれぞれ BMP 形式と PNG 形式で格納されている。

紙カードは 10 進 4 桁の番號が振られており，それに對應する各字體（代表字形）は異體字が存在しない場合はソースを示す ASCII 3 文字からなる接頭辭に紙カードの番號を付けたものを ID とし，異體字が存在する場合にはそれにさらに a, b, . . . といった接尾辭を付けたものを ID とすることで兩者の關係が紐付けられている。各字種・字體は大字典番號 [30]，および，それを基に擴張した「統合 ID」によって管理されており，これらをキーにして UCS の符號位置や大漢和番號と紐づけられている。

後に，北海道大學文學研究科池田證壽研究室に殘されていた 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數の時期に複數の作業者が作成したと思しき，重複や異同があったりデータ形式の變更や畫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。

前述のように，HNG では資料に對して番號と ASCII 3 文字による略稱（他にも，文獻名

と人間向けの略稱が存在する）が付與されているが，この番號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變化していた。また，これとは別に各資料のフォルダー名の接頭辭として使われている番號もありこれはおそらくある時期のExcelファイルに對應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號という2系統の番號に分化しているらしいということが判明した。一方，ASCII 3文字による略稱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，各版の比較檢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簡單ではなかったため，當初，48資料版のGitリポジトリをベースに作業を行い，後に各版の比較を行うことにより殘りのデータの追加を行っている。

2.2 字形整理上の問題

HNGは，版本だけでなく，手書きの寫本や拓本も收錄しているが，［14］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體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體とすると意味もなく異體字が爆發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點や線が缺けてしまったり（圖1）餘計なゴミが寫ってしまう場合（圖2）があるが，それに似た字體が存在する場合もあり（圖3），こうしたものも機械的に別字體とする（逆に，機械的に同字體と見なして統合する）のは問題であるといえる。また，闕畫をどう扱うかという問題もある（圖4）。HNGではこうした問題に關して，石塚晴通氏らの經驗や研究の蓄積に基づいた判斷が行われている。しかしながら，そうした判斷規準自體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斷することは難しい。そういう意味では，HNGのデータから推測される判斷規準を勘案して判斷する必要があるといえる。

圖1　拓本の例（開成石經孝經 0011「位」）

圖2　拓本の例（開成石經周易 0001「丹」）

圖 3　日本書紀巻二十四（圖書寮本）0002「丹」

圖 4　拓本の例（開成石經孝經 0257「世（梓）」）

2.3　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關する情報のことを「漢字構造情報」と呼ぶことにする。漢字構造情報の機械可讀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，[27] Ideographic Description Sequence（IDS）形式が ISO/IEC 10646［4］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，部品として複數の異なる包攝粒度を持つものを用いれば，複數の部品の組合せで構成される漢字の各部品の包攝範圍を示すことで，その漢字の包攝範圍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（圖 5）。[19]

多粒度漢字構造モデルにおいて，どのような包攝粒度階層を用いるかは隨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－抽象字體－抽象字形－字形という 5 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體粒度の間に統合字體粒度，抽象字體粒度と抽象字形粒度の間に詳細字體粒度を置くことを許している。

本稿では，包攝粒度付き文字情報を，超抽象文字は「〈*字*〉」，抽象文字は「〈字〉」，統合字體は「⁞字⁞」，抽象字體は「字」，抽象字形は「《字》」，字形は「『字』」のように表記することにする。

2.4　HNG 例示字形の 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攝することができなかった場合，包攝可能な抽象字體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，既存の抽象字體オブジェクトのいずれにおいても包攝することができなかった場合，包攝可能な統合字體オブジェクトの直下，もしくは，新たに抽象字體オブジェクト（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，超抽象文字と包攝粒度を上げて行き，どの包攝粒度でも包攝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攝可能な場合，漢字構造情報（IDS）を新たに記述する必要がない。

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 では各ソースに對し，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辭 ===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。

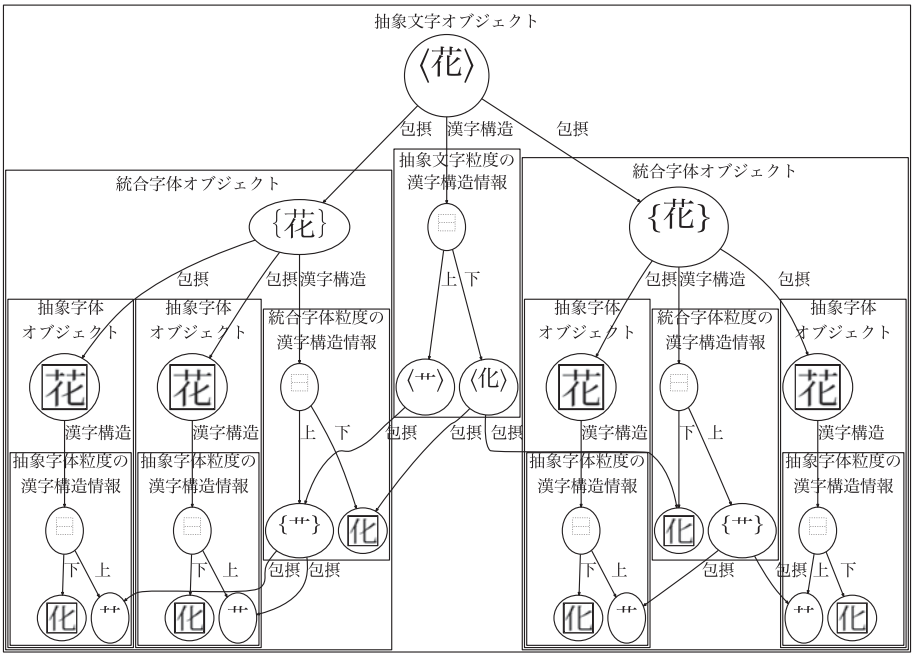

圖 5　多粒度漢字構造モデルの概念圖（花）

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，カード番號を 10 倍し，接尾辭がないものは 0，接尾辭が a のものは 1，接尾辭が b のものは 2，以下，接尾辭に對應した番號を足した番號を素性値として用いることにする。

2.5 包攝規準の問題

現在，CHISE project では，字體・字形粒度の包攝範圍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體・字形粒度の情報記述に關するガイドライン（CHISE Guidelines for Glyph Granularity of Chinese characters ; CHISE-GGG）Ver. 0.9」[23] 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，統合字體，抽象字體，詳細字體，抽象字形の包攝範圍を判定することにする。

また，抽象文字の包攝範圍は，原則として，UCS の統合漢字の符號化作業で用いられている IRG Working Document Series（IWDS）[6] 1 : List of UCV（Unifiable Component Variations）of Ideograhs（IWDS-1）を用いることにする。

抽象文字の包攝範圍をどうとらえるかに關しては，IWDS-1 に基づくものよりも廣くとらえる考え方もありうる。

例えば，IVS（Ideographic Variation Sequence）[4] は UCS 統合漢字で包攝された複數のグリフ（字體・抽象字形等）を異體字の種類を示すための枝番である VS（Variation Selector）を付けることによって區別するための仕組みであるが，このことから IVS の基底文字である統合漢字（親字）はその IVS で示されるグリフを包攝することが期待される。しかしながら，實際には IVS の登記簿である IVD（Ideographic Variation Database）[5] に登錄されているグリフの中にはその親字が IWDS-1 的には包攝可能とはいえない例が存在する。つまり，こうした IVS やグリフコレクションでは IWDS-1 よりも廣い包攝範圍を想定していると見なすことができる。

また，ISO/IEC 10646 や Unicode の規格票における統合漢字の例示字形の中にも IWDS-1 では包攝可能とはいえない例が存在する。IWDS-1 は UCS 統合漢字での包攝例や互換漢字と統合漢字の對應關係などに基づいて歸納的に定義されたものなので，UCS 統合漢字の例示字形で示されている場合はたとえ IWDS-1 では包攝可能と判斷できない場合でも包攝できるものと考えざるを得ない。

このようなケースと同樣に，HNG のデータ中には IWDS-1 では明示されていないが，抽象文字として包攝した方が良いと思われるケースが存在し，こうした場合に當初はそ

の統合漢字の包攝範圍を擴張し（包攝規準を追加し），IVS で表現されるグリフは全て包攝できるものと看做すというアプローチを採っていたが，そうした事例が發見されるたびに統合漢字の包攝範圍や包攝規準をいじると同様の部品を持つ他の漢字の構造記述や包攝範圍に影響してしまい問題が多いことが判明した。このため，現在では，UCS 統合漢字において包攝例があるもの（互換漢字として UCS 統合漢字への對應關係が定義されているもの）と IWDS-1 で包攝可能なもののみを UCS 統合漢字の包攝範圍と見なし，それ以外については筆法的變形（文字のくずし等）によって生じたような字源的に同一と思われるものをまとめた =>ucs@cognate と，複數の字源のものが衝突している可能性があるが部品としては交換可能性があるものを =>ucs@comp という擴張包攝範圍を示す ID 素性で示すことにしている。また，IWDS-1 的には包攝可能なものがソースコードセパレーション等で分離されている場合に，これらをまとめた包攝範圍を =>ucs@iwds-1 という ID 素性で示すようにしている。

2.6 簡易的な收錄

現在，初唐標準字體の例として今西本妙法蓮華經卷五の 645 字體と守屋本妙法蓮華經卷三の 592 字體，開成石經規範字體の例として開成石經論語の 1332 字體それぞれの代表字形の CHISE 文字オントロジーへの取り込み作業を完了した。

しかしながら，CHISE の包攝ポリシーに則った HNG の收錄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，勞力をかけずに既存の HNG の情報を CHISE に取り込むために，HNG の情報から機械的に變換可能な部分だけを使って文字定義を行い，HNG 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 字形オブジェクトに對して關係素性 -> HNG を張ることにした。これにより，未整理の HNG 字形もとりあえず CHISE-wiki で表示させることができ，CHISE IDS 漢字檢索の恩惠もある程度利用可能になった。

こうした簡易的な收錄を行った場合，CHISE IDS 漢字檢索の檢索對象は HNG の見出し字のものに限定され，見出し字と異なる字體の漢字構造は檢索できない。逆にいえば，今西本妙法蓮華經卷五と守屋本妙法蓮華經卷三と開成石經論語では見出し字と異なる字體の漢字構造情報も機械可讀化されているため，例えば，「十」と「刀」を部品として持つ漢字を檢索することで「  」を探すことができる。

こうしたことを鑑みれば，宮廷寫經や開成石經といった各時代の標準字體を代表するような資料に對して CHISE の包攝ポリシーに則った收錄作業を行うことで，HNG の全ソースの收錄作業を行うことなく，一定の利便性を確保できるのではないかと思われる。

2.7　CHISE-IDS HNG 漢字檢索

「CHISE-IDS HNG 漢字檢索[5)]」は「CHISE IDS 漢字檢索」の檢索對象を HNG の範圍に限定したものである。CHISE IDS 漢字檢索と同樣に探したい漢字に含まれる部品を檢索窓に入力すると指定した部品を含む漢字の一覽が表示される（圖 6)。また，CHISE IDS 漢字檢索と同樣に檢索結果の左端の文字畫像をクリックすると文字の詳細畫面が表示される。

このように CHISE-IDS HNG 漢字檢索は CHISE IDS 漢字檢索と基本的に同一のサービスであるが，檢索對象を HNG の範圍に限定することにより，HNG 用例のある文字だけを容易に檢索することができる。これにより，共通する部品を持つ複數の漢字の字體用例を簡單に比較することが可能になった。

圖 6　CHISE-IDS HNG 漢字檢索

3　カード畫像のサポート

HNG 公開 10 周年イベントの場で，ユーザーから「石塚漢字字體資料」の紙カード畫像を公開して欲しいという要望があり，石塚晴通氏他關係者が一般公開することで合意

5）　http : //www.chise.org/hng-ids-find

したため，2016 年 4 月に CHISE-wiki に對して石塚漢字字體資料の紙カード畫像の表示機能を追加した。また，2016 年 6 月には IIIF Image API を利用した紙カード畫像と開成石經拓本畫像の配信，及び，これを利用した京大人文研所藏の開成石經畫像と HNG の開成石經データの比較表示機能も追加した。

3.1 HNG の紙カード畫像

「石塚漢字字體資料」は，漢字の字體には時代・地域（國）による標準が存在し，その標準は時代・地域により變遷することを示すために，漢字文化圈の各時代の標準的な文獻（漢籍・佛典・國書等の典籍）を中心に，比較のため，私的な文獻も交えて選定し，漢字字體の用例調査を行い，紙カードとして蓄積していたものである。

この用例調査では，選定された文獻（以下では「ソース」と呼ぶことにする）毎に，そのソースに出現する字形を字種毎にまとめて，1 つの紙カード（圖 7）にまとめる形で行われていた（字種は「大字典」に基づいて整理されていた）。但し，ある字種に屬する字形の數が多すぎて 1 枚の紙カードに張り付けられない場合は複數のカードを用いている（圖 8）。ある字種に複數の字體が認められる場合も 1 つ（もしくは一連）の字種毎の紙カードにまとめられており，字體の數が紙カード上部の三角の切込み位置で表現されている。

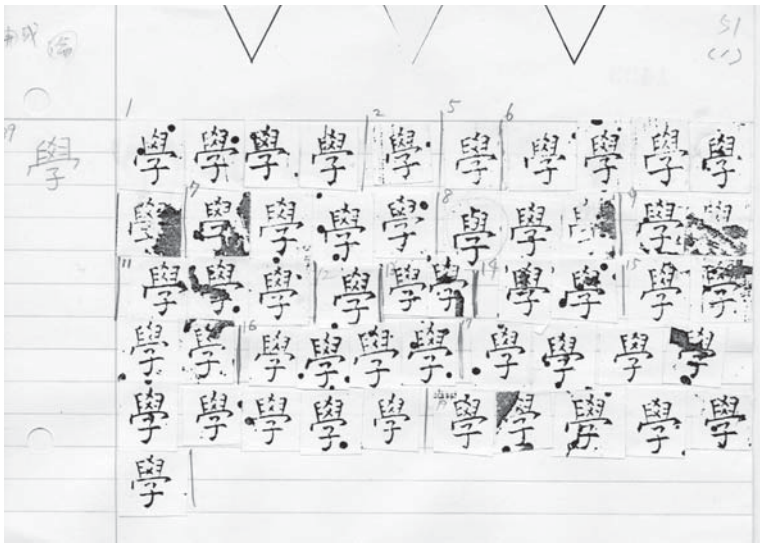

圖 7　紙カードの例（開成石經論語―學）

HNG はこの紙カードの情報を電子化したものを基礎に構成されている。

HNG ではこの紙カードをスキャンしソース毎に 10 進 4 桁の番號を付した畫像ファイルにしている。

字種もまたこのソース毎に 10 進 4 桁の番號で管理されている。卽ち，ある字種の紙カードが 1 枚の場合はその紙カードの畫像ファイルの番號が字種の番號となる。もし，ある字種の字形用例が複數の紙カードで構成されている場合は，先頭の紙カードの畫像ファイルの番號がその字種の番號となる。そして，この字種の番號とソースの ID との對によってユニークな ID を構成することができる。

字體は字種の ID に對して，a, b, . . . という枝番を付與することで識別している。

このように，HNG では字種や字體の ID は紙カードの番號に基づいて生成されており，字體や字種と紙カードは紐づけられている譯である。

3.2 設計

從來，CHISE と HNG の統合は HNG における代表字形を CHISE の字形粒度の文字オブジェクトとして表現し，CHISE の文字オブジェクトの世界に取り込むことで實現していた。すなわち，『文字』の世界，CHISE が用いている有向グラフ型データベースエンジン **Concord**［18］の用語でいえば，character ジャンルだけを使って表現していた。

HNG の代表字形に「石塚漢字字體資料」の紙カードの畫像を紐づけることを考えた場合，最も單純なのは，紙カードの畫像をどこかの Web サーバーに置いておき，HNG 字形オブジェクトに對應する紙カードの畫像をリンクすることである。

しかしながら，紙カードに對してメタデータを記述したり，なんらかの言及（アノテーション）を行ったり，別の何かと關連づけようとする場合，紙カードを單なる畫像ファイルとするだけでは不十分であり，Linked Data におけるノードとして扱うのが望ましいといえる。つまり，紙カードの畫像とは別に抽象的な概念ないしはものとしての紙カードを示すオブジェクトを想定するのが望ましい。

よって，HNG 字形オブジェクトと HNG 紙カードオブジェクトと HNG 紙カード畫像オブジェクトという 3 種類のオブジェクトを設け，HNG 字形オブジェクトと HNG 紙カードオブジェクト，HNG 紙カードオブジェクトと HNG 紙カード畫像オブジェクトをリンクすることにする。

3.3 實装

3.3.1 畫像リソースと畫像配信サービス

一般に，ある畫像ファイルを Web 上で配信する方法は複數ありうる。同じファイルを複數の異なるサーバーに置く場合もありうるし，畫像形式を變換する必要がある場合もある。また，ファイルサイズの大きな高精細畫像を配信する場合，複數解像度タイル畫像として扱った方が良いかもしれない。現在，デジタル人文學の世界を中心として，こ

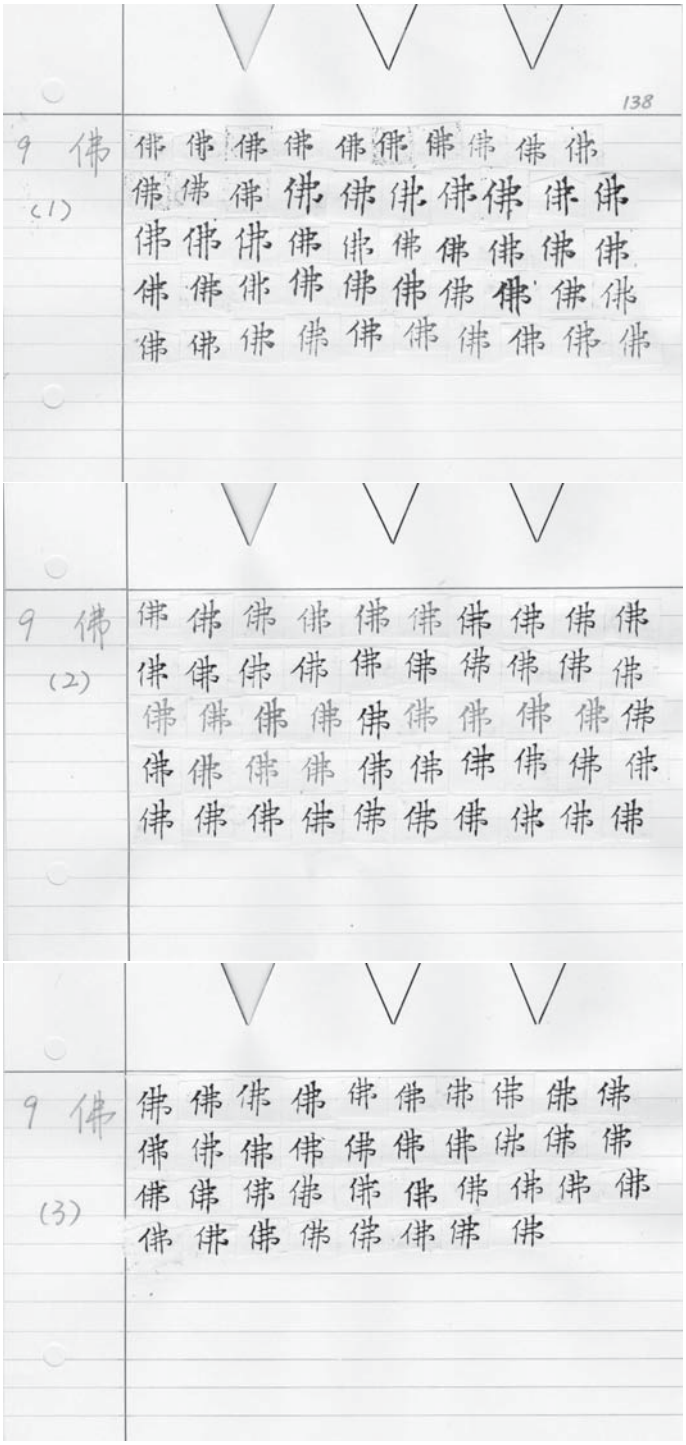

圖 8　複數の紙カードの例（守屋本妙法蓮華經卷三 ― 佛）

のような畫像配信用 Web API の仕様として IIIF Image API［1］が普及しつつあり，相互接續性を考慮した場合，このような標準的な API を用いることも考慮すべきであるといえる。しかしながら，Web API は Web の世界の變容とともに變わっていく可能性も大いにありうるし，新たな畫像配信の仕組みや形式に對應する必要に迫られる場合もあるかもしれない。

こうしたことを鑑み，畫像ファイルとその配信サービスを區別したモデル化を行い，抽象的な畫像ファイルを示すオブジェクトの屬性としてその內容を配信可能なサービスを列擧するという方法を採ることにした。

畫像配信サービスの實裝としては，JPEG ファイルをそのまま配信する方法に加えて，IIIF Image API をサポートするために IIPImage［3］server Version 1.0 を採用した。IIPImage server は，その名の通り，IIP（Internet Imaging Protocol）に基づいて畫像を配信するサーバーである。IIPImage server は IIP の獨自擴張の一種として IIIF Image API をサポートしており，1 つのサーバーで IIP と IIIF Image API の兩方をサポートすることができる[6)]。よって，IIIF Image API だけでなく IIP もサポートすることにした。

3.3.2 畫像リソースオブジェクト

ある畫像ファイルを抽象化したものとして「畫像リソース」オブジェクトを設けた。これは Concord の image-resource ジャンルのオブジェクトとして實現している。

このオブジェクトは

=id 素性　ID を示す。シンボル型。必須。
=location 素性　畫像資源の識別子（URL）を示す。文字列型。必須。
=location@iiif 素性　IIIF Image API での URL を示す。文字列型。オプション。
=location@iip 素性　IIP での URL を示す。文字列型。オプション。
image-width 素性　畫像の幅（ピクセル數）を示す。整數型。必須。
Image-height 素性　畫像の高さ（ピクセル數）を示す。整數型。必須。
name 素性　畫像資源の名前を示す。文字列型。必須。
<-image-resource 素性　この畫像を表示とするオブジェクトを示す。オブジェクトのリスト型。オプション。

といった素性を持つ。

6） IIP と IIIF Image API を混在して使用することもできる。

3.3.3 HNG カードオブジェクト

「石塚漢字字體資料」の紙カードを表現したものとして「HNG カード」オブジェクトを設けた。これは Concord の hng-card ジャンルのオブジェクトとして實現している。

このオブジェクトは

=id 素性　ID を示す。シンボル型。必須。
=hng-id 素性　ID を示す。シンボル型。必須。
name 素性　紙カードの名前を示す。文字列型。必須。
representative-glyph 素性　對應する HNG 代表字形オブジェクトを示す。オブジェクトのリスト型。オプション。
->image-resource 素性　紙カードを寫した寫眞の畫像リソースを示す。オブジェクトのリスト型。必須。

といった素性を持つ。

=id 素性と=hng-id 素性の値は 18-1234 のようにソース番號とカード番號の 10 進數を '-' で繋げたものを用いた。

3.3.4 HNG 代表字形オブジェクト

從來から CHISE 文字オントロジー中に設けていた HNG 代表字形オブジェクト（Concord の character ジャンルのオブジェクト（文字オブジェクト））に

sources@HNG/card 素性　對應する HNG カードオブジェクトを示す。オブジェクトのリスト型。

を追加した。

HNG 代表字形の ID と紙カードの ID は元々同じものであり，機械的に變換可能であるので，この情報は冗長であるが，CHISE-wiki（EST [21]）は Concord 中に存在する Linked Data のグラフに書かれたことしか知らないため，特別な處理を行わなければ紙カードの畫像を表示したりリンクすることができない。しかしながら，このような冗長な情報をあらかじめ靜的に生成しておくことはメンテナンスコストをあげるため，HNG 代表字形オブジェクトの CHISE-wiki 頁（例：http : //www.chise.org/est/view/character/repi.hng-kar=12340）にアクセスした際に動的に生成するようにしている。

4　拓本文字データベースの統合

HNG はさまざまな字形用例のデータを収錄しているが，「石塚漢字字體資料」の紙カードをベースにしているという技術的な理由と，利用許諾の問題という 2 つの制約のために，全文畫像を表示することができなかった。

一方，「拓本文字データベース」では京都大學人文科學研究所所藏の拓本畫像に座標データ付の文字情報が埋め込まれており，字形用例から元の拓本畫像を表示可能である。しかしながら，拓本文字データベースでは字形用例の整理がなされておらず，HNG のように字體毎に整理して表示したり分析することができない。

こうしたことから，兩者を繫げることで HNG に對して全文畫像表示機能を提供するとともに，拓本文字データベースに對して HNG へのマッピングを提供することを目指した。

こうしたことを實現するためには HNG と拓本文字データベースの兩者で共通したソースが存在しなければならない。そして，それに該當する可能性があるのが開成石經の孝經と論語，周易であったので，これらを對象に檢討を行った。

HNG と「拓本文字データベース」は共に開成石經のデータを持っているが，元になっている拓本は異なっており，前者が東洋文庫所藏の拓本を用いているのに對し，後者は京都大學人文科學研究所所藏の拓本を用いている。よって，嚴密にいえば，拓本上の字形は同一とはいえないが，同じものの拓本には違いがなく，そう大きな差異はないであろうと假定してこの兩者を對應づけることにした。

4.1　拓本文字データの表現

4.1.1　元データ

拓本文字データベースでは各拓本に含まれる文字の情報が畫像マークアップされている。拓本中の各文字は拓本畫像の座標を使ってその領域が示されるとともに，その文字の UCS での符號位置が記されている。

4.1.2　拓本畫像の畫像リソースと配信用サービス

拓本畫像の配信用サービスとしては，紙カード畫像と同樣に，IIPImage server を用いることにした。そして，畫像リソースオブジェクトによって拓本畫像を表現することにした。但し，拓本畫像では，=location@iiif 素性と=location@iip 素性に加えて，=location@djvu 素性を用いて拓本文字データベースでの URL も示すようにした。

また，4.1.3 節で述べる<-image-segment 素性の逆關係素性 ->image-segment と，4.1.4 節で述べる <-segmented-glyph-image 素性の逆關係素性 ->segmented-glyph-image が生成される。前者はある拓本畫像に含まれる部分畫像（文字畫像）の一覽を示し，後者

は出現字形の一覧を示す。

圖 9　拓本畫像リソースの EsT での表示例（開成石經周易 2）

圖 9 に畫像リソースオブジェクトの EsT での表示例を示す。ここでは ->segmented-glyph-image 素性の値である出現字形オブジェクトをそれに對應する文字として表示しているため，ブラウザーの檢索機能を使ってテキストに含まれる文字列を檢索することができる。この例では ->segmented-glyph-image 素性の値が省略されているが，「…」をクリックすれば省略されている殘りの部分も表示できる。

また，畫像リソースオブジェクトが =location@iiif 素性を持っている場合，EsT の頁上部の主題表示部に OpenSeadragon［9］を使って =location@iiif 素性の値で示される畫像を表示するようにした。このため，畫像をドラッグして表示位置を變えたり，OpenSeadragon のアイコンを押して擴大縮小したり全畫面表示することも可能である。

4.1.3　部分畫像オブジェクト

出現字形の畫像を表現するために，元となる畫像の中の矩形領域で示される部分畫像を表現するための Concord オブジェクトを設けた。これを「部分畫像オブジェクト」と呼ぶことにする。

部分畫像オブジェクトは，HNG の紙カード畫像や拓本畫像と同樣に，image-resource ジャンルのオブジェクトとするが，全體畫像の場合に加えて，

image-offset-x 素性　全體畫像中での部分領域の始點の X 座標を示す。整數型。必須。

image-offset-y 素性　全體畫像中での部分領域の始點の Y 座標を示す。整數型。必須。

<-image-segment 素性　この部分畫像オブジェクトに對應する全體畫像オブジェクトを示す。オブジェクトのリスト型。必須。

<-image-resource 素性　この畫像を表示とするオブジェクトを示す。オブジェクトのリスト型。オプション。

=location@djvuchar 素性拓本文字データベースでの部分畫像の URL. 文字列型。必須。

という素性を付與することにした。

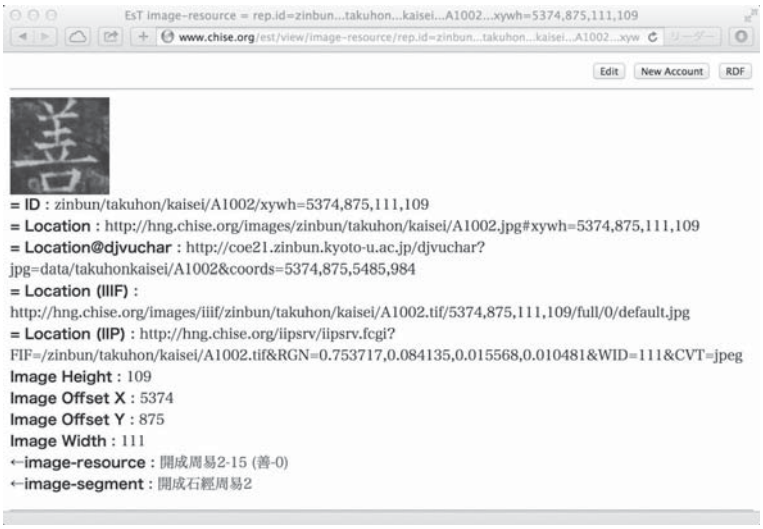

圖 10　部分畫像の EsT での表示例

圖 10 に部分畫像オブジェクトの EsT での表示例を示す。ここで，先頭に表示されている部分畫像をクリックすると全體畫像での出現位置にジャンプすることができる。また，<-image-resource 素性の値部をクリックすると，4.1.4 節で述べる出現字形オブジェクトに飛ぶことができる。

4.1.4　出現字形オブジェクト

拓本中の各文字（字形）を表現するために，出現字形オブジェクトを設ける。これは HNG 字形オブジェクトとは異なり，文字オブジェクト（character ジャンルのオブジェクト）とはせず，新たに glyph-image ジャンルの Concord オブジェクトを設けることにした。このオブジェクトは

=id 素性　ID を示す。シンボル型。必須。

character 素性　出現字形に對應する UCS の抽象文字を示す。オブジェクトのリスト型。オプション。

name 素性　出現字形の名前を示す。文字列型。オプション。

<-glyph-image@zinbun/takuhon 素性　對應する HNG カードオブジェクトを示す。オブジェクトのリスト型。オプション。

<-segmented-glyph-image 素性　この字形が出現した拓本畫像リソースオブジェクトを示す。オブジェクトのリスト型。必須。

-> image-resource 素性　この出現字形の畫像リソースを示す。オブジェクトのリスト型。必須。

といった素性を持つ。

圖 11　出現字形の EsT での表示例

圖 11 に出現字形オブジェクトの EsT での表示例を示す。ここでは，character 素性に「→ UCS 抽象文字」，<-glyph-image@zinbun/takuhon 素性に「→ HNG カード」という素性名の表示を設定している。また，部分畫像オブジェクトの場合と同樣に，先頭に表示されている字形畫像をクリックすると全體畫像での出現位置にジャンプすることができる。

4.2　HNG カードオブジェクトを用いた情報統合

拓本文字データベースの出現字形の情報と「石塚漢字字體資料」の字形用例の情報を統合する方法として，ある紙カードの字種に相當する拓本文字データベースの出現字形オブジェクトの集合をその HNG カードオブジェクトの中の素性として表現する方法を採ることにした。そして，このための素性として，4.1.4 節で述べた <-glyph-image@zinbun/takuhon 素性の逆關係素性である ->glyph-image@zinbun/takuhon 素性を用いることにした。

これは，要するに，拓本文字データベースの開成石經の出現字形を「石塚漢字字體資料」と同様な方法で整理して假想的な紙カードを作り，それを元々の「石塚漢字字體資料」の紙カードと併置するようなものだといえる。これによって，HNG カード畫像（東洋文庫所藏の拓本字形）と拓本文字データベースの出現字形（京大人文研所藏の拓本字形）を見比べることができる譯である（圖 12）。

圖 12　HNG カードの EST での表示例

4.3　HNG 代表字形オブジェクトでの表示

3.3.4 節で述べたように，HNG カードオブジェクトをサポートするために，HNG 代表字形オブジェクトに sources@HNG/card 素性を追加したが，拓本文字データベースの出現字形の情報を統合するために

sources@zinbun/takuhon 素性　對應する拓本出現字形オブジェクトを示す。オブジェクトのリスト型。

を追加した。

sources@HNG/card 素性と同様に，この素性も HNG 代表字形オブジェクトの CHISE-wiki 頁にアクセスした際に動的に生成（HNG カードオブジェクトの ->glyph-image@zinbun/takuhon 素性の値をコピー）するようにしている。

圖 13 に HNG 代表字形の CHISE-wiki（EST）での表示例を示す。ここでは，sources@HNG/card 素性に「出典（石塚漢字字體資料）」，sources@zinbun/takuhon 素性に「出典（拓本文字データベース）」という素性名の表示を設定している。

この例では sources@zinbun/takuhon 素性の値としてこの字形とは異なる出現字形が含まれていることが判るが，これは拓本文字データベースでは出現字形の字體への分類作業が行われておらず，單純に UCS 統合漢字の抽象文字に對應づけているだけだからである[7]。

HNG カードオブジェクトを用いた情報統合の場合，「石塚漢字字體資料」の紙カードは字種單位にまとめられているので，拓本文字データベースの出現字形を UCS 統合漢字の抽象文字によって統合しても問題が生じなかったが，HNG 代表字形オブジェクトに統合する場合，このような氣持ち悪い事態に遭遇し得る譯である。

しかしながら，開成石經では 1 つの字種において異體字が使われる例は極めて少なく，實際にはこのようなケースはほとんど生じないといえる。

むしろ，現實的に HNG と拓本文字データベースの情報統合で問題となるのは，HNG の見出し字と拓本文字データベースの UCS のコードポイントが一致していないケースである。HNG では「大字典」に基づいた見出し字以外に幾つかの異體字も含めているが，必ずしも網羅的でないために，UCS が複數の異體字を收錄している場合に食い違うことがある。また，拓本文字データベースの作成過程において既存の電子テキストを利用して畫像マークアップしたために，開成石經での表現とは異なる異本の文字が埋め込まれてしまったケースもある。また，入力ミスと思われるものもあり，異體字處理だけでは解決できないといえる。

こうした不一致のケースはこれまでの所，25 件見つかっているが，こうした場合，拓本文字データベースのテキスト情報を利用して，異體字情報や論語の異同情報，前後關係の情報等を利用することで對應する出現字形を見つけることができた。また，こうしたものが見つかった場合，HNG 代表字形オブジェクトの <-HNG@zinbun/takuhon 素性の値に拓本文字データベースでの UCS の抽象文字を入れることで兩者の對應關係を表現するようにしている[8]。

7） 實際にはさらに異體字をどれかに寄せている場合がある。

8） これにより，拓本文字データベースでの UCS の抽象文字の ->HNG@zinbun/takuhon 素性の値にこの HNG 代表字形オブジェクトが入り，HNG と拓本文字データベースの兩方の見出し文字から HNG 代表字形オブジェクトに到達することができるようになる。

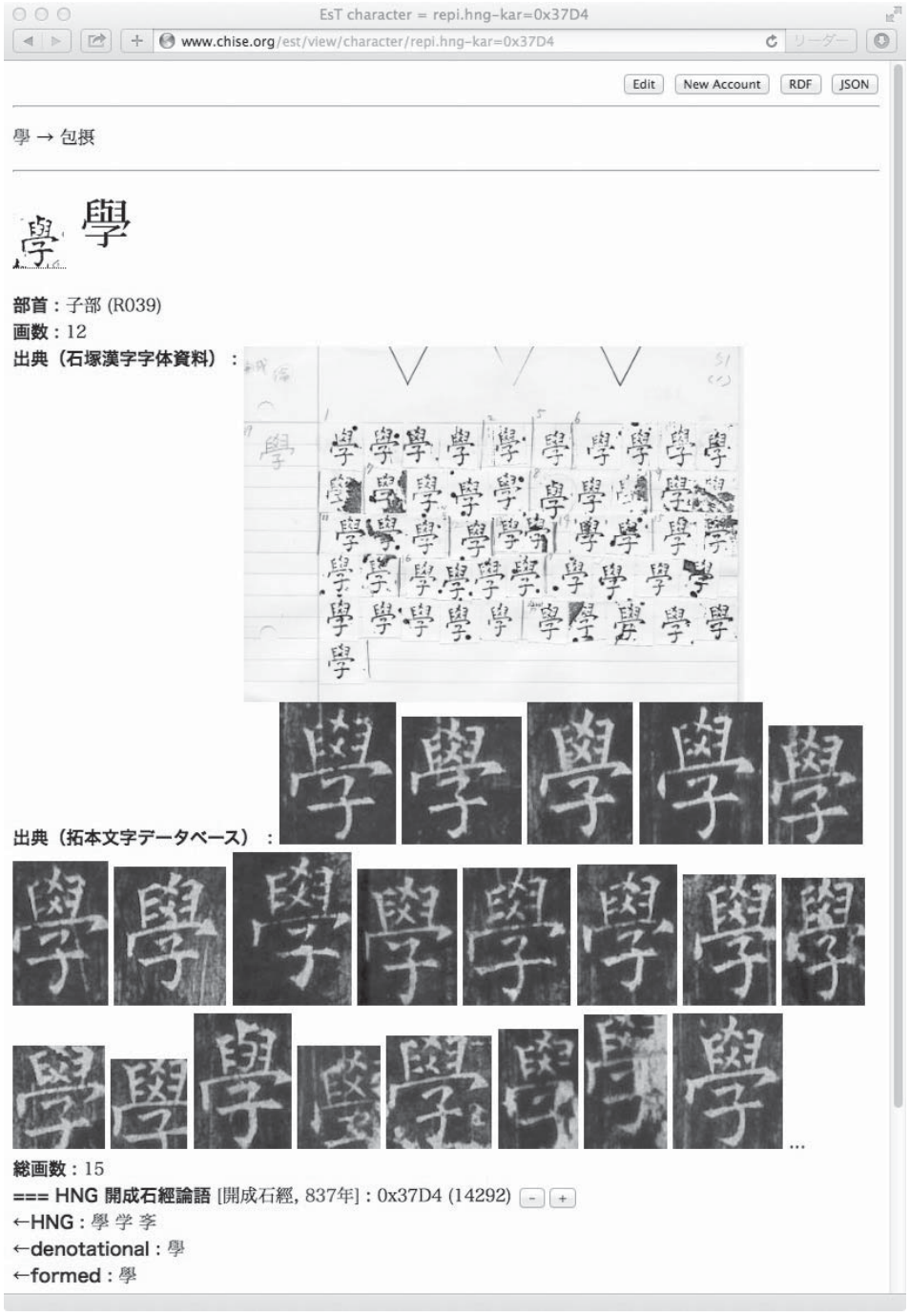

圖 13　HNG 代表字形の CHISE-wiki（EsT）での表示例

5　資料の發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，數年に一度位に，大きなアーキテクチャー變化の波がやって來ており，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續くかどうか疑わしく，デバイスの變化，ネットワークインフラの變化，ソフトウェア技術の變化，社會的な變化等を背景に今後もコンピュータ環境は變化し續けると考えられる。

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變化の波を乘り越える必要があり，それを乘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變化して行くので，設計時に普及していた製品や技術，運用體制や使われ方といったもろもろの前提が數年後には崩れてしまうということも少なくない譯である。

こうした問題に對處する方法の一つに，データベースから Web 上での檢索サービスや 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變化に合わせた今後の發展を鑑みれば，從來の HNG で公開していた漢字字形の切り抜き畫像とメタデータに加え，石塚漢字字體資料の紙カード畫像と聞き取り調査やデータ發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。「漢字字體規範史データセット」[28] 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。

5.1　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，その擧動を嚴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數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檢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關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，發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，關係者といえども全てに關っている譯ではない上，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，論文等で發表されていないことがらの意圖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（人閒向けに）文書化するとともに，機械可讀な形でも表現することに取り組んでいる。ここでの對象となる事項にはデータ形式や檢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，各資料に關する情報やその選定理由等，あるいは，「書體」や「字體」，「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搖れがあったり石塚漢字字體資料や HNG において獨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，他の概念との關係等によって（石塚晴通化が元々どういう意味づけをしていた・どういう意圖で使っていたかや石塚漢字字體資料・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，［15］や［16］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豫定である。

一方，HNG の內容やモデルを理解する上で，石塚漢字字體資料や 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，元資料に當たることも重要であるが，その解釋や分類等の HNG 固有の部分が問題となった時や，對應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體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，北海道大學文學硏究科池田證壽硏究室の協力を得て，その資料調査を行った。

5.2　データセットの保存と利活用

データベースの永續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爭的資金によって開發され，プロジェクト終了後に豫算的・人員的問題から十分な運用體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數の關係者の努力と熱意（とお金）に依存してしまい，機械の故障やセキュリティー對策，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（擔當者

が燃えつきてしまい）對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，お金がなくても持續できるような，言い替えれば，競爭的資金に依存しなくてもすむような仕組みを實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，逆說的ではあるが競爭的資金に依存しない體制を採るためには，多くの場合，競爭的資金を使ってその維持管理體制を改修する必要があると思われる。HNG の場合も同樣であり，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 による畫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體制の確立を目的とした科研費の申請をした所採擇され[9]，實際にその實現に向けて動くことができた譯である。

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體制をとることが重要であり，最惡の場合，利用者の小額の寄付によって維持できるぐらいの體制にすることが望ましいと考えられる。

一方，2 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 關連サービスを提供したが氣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 は研究者が所屬する組織の改組や研究者の移動等によって變化しないようにするべきであり，このため，hng-data.org という獨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，「漢字字體規範史データセット保存會」を設立し，その設立イベントを通じて Web サイトの廣報活動を行った。

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，獨自ドメイン上で GitLab Community Edition を用いて獨自のホスティングサービス gitlab.hng-data.org を立ち上げたのは，營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險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

9） 基盤研究（C）「字體記述の精密化手法の確立による歷史的漢字字體情報アーカイブズ構築」（18K00611）

は無料で利用できたとしても，ある時大きくポリシーが變わって不都合が生じたり，サービスが停止したり，致命的にその內容が變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觀點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體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）[2] [10] の利用も檢討している。

5.3　漢字字體規範史データセット

現在，我々は石塚漢字字體資料および漢字字體規範史データベースの內容を長期にわたり安定的に利用可能にするために，これらの情報や關連する情報を調査・整理し，Git によって版管理されたデータセットにまとめる取組みを行っている。これを「漢字字體規範史データセット」[17]（「HNG データセット」と略す）と呼ぶ。

このデータセットの Git リポジトリ [28] は前述の CHISE を利用した HNG 關連サービスの開發時に作成したものをベースにしている。これは 2006 年度時點でのバックアップデータをベースに開發が始まったため，48 資料しか收錄できていなかった [8] が，最終的に公開上の問題がないと考えられる 63 資料を收錄する豫定である。

2019 年 4 月現在，60 資料を收錄しており，最新版だけでなく，過去の版の情報もブランチやタグ等を利用して參照可能にすることを計畫している。

6　從來型檢索 UI の再現

CHISE-wiki 上での HNG 字形表示機能と紙カード畫像表示機能，および，「CHISE-IDS HNG 漢字檢索」を用いた部品による HNG 漢字檢索機能により Web 上で HNG のデータを檢索・閱覽することができるようになった。こうした CHISE と HNG の統合により，部品單位での檢索や比較，紙カード畫像に戻っての再檢討，HNG の開成石經の拓本よりも良い拓とされる京大人文研所藏の開成石經畫像との連携機能といった從來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，舊來のユーザーインターフェース（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 [26]。

そこで，我々は停止前の「漢字字體規範史データベース」（舊 HNG 檢索）に近い操作性や畫面表示を持ちつつ，現在の Web 環境に適合し，また，CHISE との統合によってもた

らされた新たな機能も利用可能な漢字字體規範史データセット用の檢索サービスとして「漢字字體規範史データセット單字檢索」（略稱「HNG 單字檢索」）

https : //search.hng-data.org/

を開發した。

6.1 スクリーンショットの收集と分析

このシステムの開發計畫が始まった 2018 年秋頃にはオリジナルの HNG が停止して既に 3 年餘りが經っており，その擧動に關する記憶は徐々に風化しつつあった。このため，その仕様を確定することは，當初の想定に反し，實の所容易なことではなかった。また，使用頻度の低い機能や 6.2.1 節で述べるような今日の Web 環境に合わないような部分なども含めてオリジナルの HNG を完全再現することが良いとは必ずしもいえず，そのエッセンスを殘しつつ妥當で實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。これは論文や學會發表等でのスライド，記事等で引用される形で斷片的に殘っているが，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・どの版のものを參照しているかが判りづらいことが少なくない。HNG は何回かバージョンアップを重ねており檢索結果も變化しているため，各版・時期における畫面表示がどのようなものであったか（いつどのように變化していったか）が判ると良いのであるが，このために必要となる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狀は非常に嚴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 サービスの將來に向けての保存という觀點では，スクリーンショットの網羅的保存は重要な檢討項目ということができるだろう。そういう譯で，我々は極めて限られたスクリーンショットとうろ覺えの記憶を賴りに畫面設計を行わざるを得なかった。

オリジナルの HNG の檢索結果の表示畫面では，上から順に中國寫本，中國版本（石經を含む），日本資料，その他という順に多段表示されていた（圖 14）が，HNG データセットの各資料に付けられた「初唐寫本」や「日本版本」といった資料の分類に關する項目（[16] では「區分」としている）はこの機能を實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「大和寧寫本」や「北宋版」などのように，中國寫本，中國版本，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

圖 14　オリジナル HNG での表示例

圖 15　従來の CHISE-wiki での表示例

見出し文字と HNG 字形の對應關係を HNG の見出し文字に對應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對して關係素性 ->HNG を張ることで表現していた。[24] [25] CHISE-wiki ではこの關係素性 ->HNG をその標準機能によって單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譯である（圖 15)。

CHISE-wiki において，オリジナルの HNG 風の檢索結果を實現する場合，全てを關係素性 ->HNG にまとめるよりも各資料の區分に基づいたドメインに分けた方が良いといえる。假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單だと考えられる。そこで，中國寫本，中國版本（石經を含む)，日本寫本，日本版本，韓國資料，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關係素性で表現することにした（圖 16)。表 1 に HNG データセットにおける資料の區分との對應を示す。なお，從來，日本資料としてまとめられていたものを日本寫本と日本版本の 2 つに分けたのは HNG が設計された頃と比べその後の增補により日本資料の數が增えたことと開成石經規範の影響に寫本と版本で差があるケースを考慮したためである。

この際，具體的に，どの資料（區分）をどの種別に對應させるかが問題であることが判

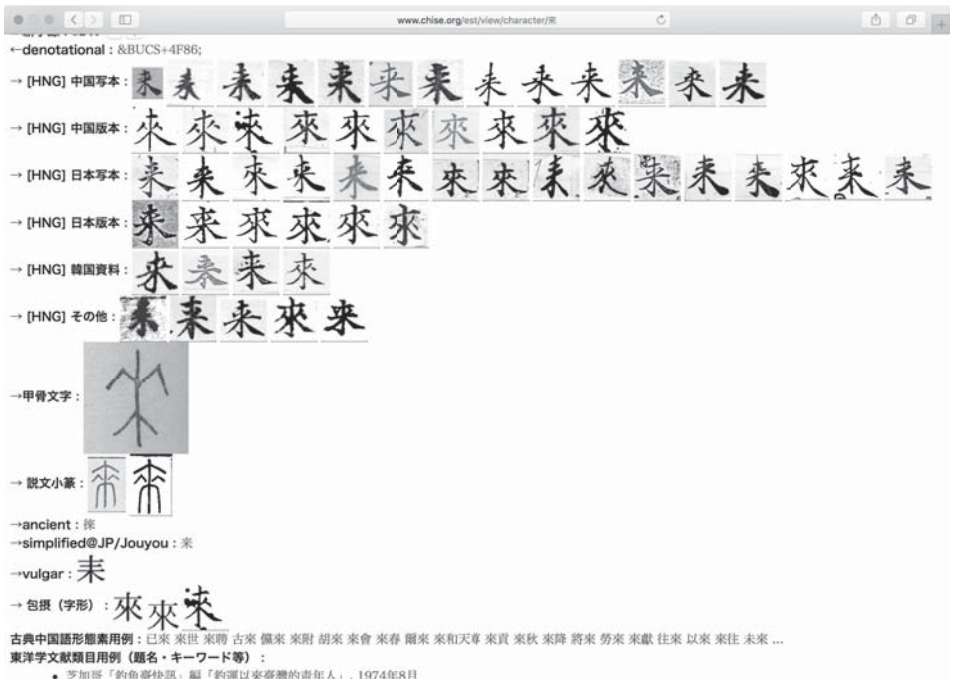

圖 16　現在の CHISE-wiki での表示例

り，この件について調査を行い，對應關係の確定を行うこととなった。

このように，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。卽ち，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，檢索や表示といった檢索サービス上でのなんらかの擧動を實現するために設けられている譯であり，そうした擧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，檢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處理するかを示すことはその仕様の一部であろう。

もし，關數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檢索を實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，現實の關數型言語や論理型・制約型言語で書かれたプログラムは現實の實裝を反映し，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實現することができれば，プログラム，卽ち，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區分と檢索結果の表示畫面での多段表示の關係は表 1 に示すような各資料の區分と段の寫像として表現できるし，同一資料に複數字體が存在する場合の各例示字形の配置法は「字體數」の項目を使って定義できる。もちろん，實際のシステムにはより低レベルの部分が含まれるが，HNG としての意圖を形式的

表 1　資料の區分とドメインの對應

區　分	ドメイン	種　別
南北朝寫本	CN/manuscript	中國寫本
隋寫本	CN/manuscript	中國寫本
初唐寫本	CN/manuscript	中國寫本
則天寫本	CN/manuscript	中國寫本
盛唐寫本	CN/manuscript	中國寫本
高昌寫本	MISC	そ の 他
吐蕃寫本	MISC	そ の 他
大和寧寫本	MISC	そ の 他
開成石經	CN/printed	中國版本
北宋版	CN/printed	中國版本
南宋版	CN/printed	中國版本
西夏版	MISC	そ の 他
日本寫本	JP/manuscript	日本寫本
日本版本	JP/printed	日本版本
日本書紀寫本	JP/manuscript	日本寫本
韓國寫本	KR	韓國資料
韓國印刻本	KR	韓國資料

に示す上でそうした詳細な實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當程度宣言的に記述可能であると考えられる。

6.2 要求仕様

6.2.1 システムに關するもの

舊來風の UI を持った檢索サービスを開發するにあたり，單に從來風の操作性や表示を再現するだけではなく，現代的な Web 技術を用い現在の Web 環境と調和した實装を實現することを目指した。

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內容を全部取って行くことを懸念して，わざと檢索結果や畫像の 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檢索結果に對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觀點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實現したいと考えた。

また，CHISE と連係することで，紙カード畫像や拓本畫像の利用や CHISE の文字情報，Unihan/GlyphWiki 等との連携といった CHISE-wiki（EsT［21］）上で實現された新たな機能を利用することも目指した。

こうした觀點に基づき，

1. 檢索結果に對するパーマリンクの實現
2. レスポンシブ Web デザインの實現
3. 石塚漢字字體資料の紙カード畫像（EsT）へのリンク
4. CHISE-wiki 上の HNG 例示字形オブジェクトの頁へのリンク
5. 見出し字（UCS 抽象文字），大字典番號（字種），大漢和番號（例示字體）に對應する CHISE-wiki 上の文字オブジェクトの頁へのリンク

を要求仕様とした。

6.2.2 データに關するもの

從來の CHISE ベースの HNG 關連サービスは 48 資料のみをサポートしていたが，HNG 單字檢索では HNG データセットが收錄對象としている 63 資料全てを對象とすることにした。

また，舊 HNG 檢索では見出し字および異體字の符號化において BMP の範圍內でしか

UCS 符號化を行っていなかったが，漢字字體規範史データセットでは統合漢字擴張 F までの現時點で利用可能な UCS に收錄された統合漢字の全てをサポートすることを目指し，統合漢字擴張 B 以降へのマッピング情報の追加を目指すことにした。これに伴い，HNG 單字檢索も舊 HNG 檢索では利用できなかった擴張 B 以降の統合漢字をサポートする。

6.3 實裝

6.3.1 システム

HNG 單字檢索のサーバー側の實裝には PHP で書かれたオープンソースの Web アプリケーションフレームワークである Laravel を用いた。また，フロントエンドにはオープンソースの JavaScript フレームワークである Vue.js 及び Vue.js 用のマテリアルデザインフレームワークである Vuetify を用いた。圖 17 にその構成圖を示す。

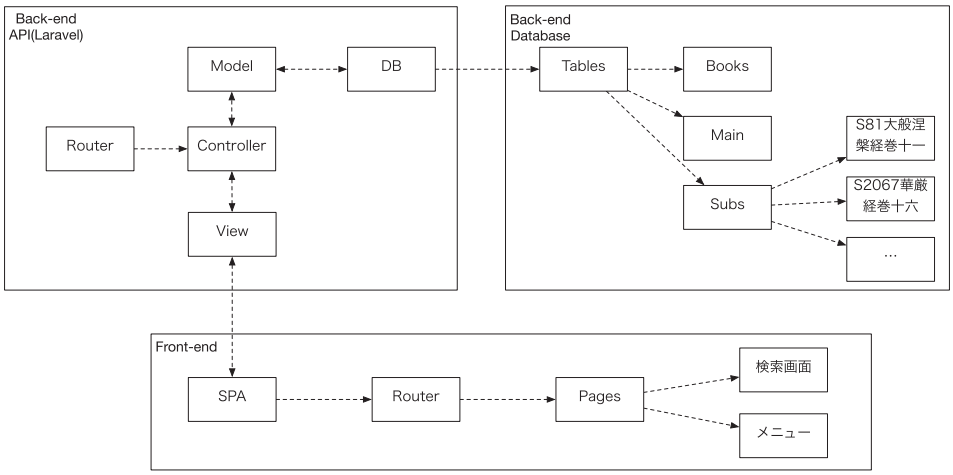

圖 17　HNG 單字檢索の構成

6.4 データセット

2019 年 4 月現在ではまだ HNG データセットに收錄されていないものも含め 63 資料を先行して收錄した。

また，CHISE ベースの HNG 關連サービスにおいても資料の追加作業を行っており，2019 年 4 月現在，60 資料を收錄している。この 60 資料に關しては紙カードの畫像も表示可能である。

6.5 Web UI

「漢字字體規範史データセット單字檢索」（略稱「HNG 單字檢索」）のトップページである

https : //search.hng-data.org/

にアクセスすると圖 18 のような畫面が現れる。HNG 單字檢索の使い方は基本的には說明文の下もしくはヘッダー部にある檢索窓（どちらでも良い）に探したい 1 文字を入力するだけという極めて單純なものである。

圖 18　HNG 單字檢索のトップページ

6.5.1　檢索結果の URL

例えば，檢索窓に「學」を入れて檢索ボタンを押すと圖 19 のような檢索結果を表示する畫面が現れ HNG の各用例が表示される。

この畫面は指定した檢索對象である文字を c とする時，これに對應した固有の URL

https : //search.hng-data.org/search/c

を持つ。

例えば，「學」を指定した場合，その檢索結果のパーマリンクは

https : //search.hng-data.org/search/學

となる。

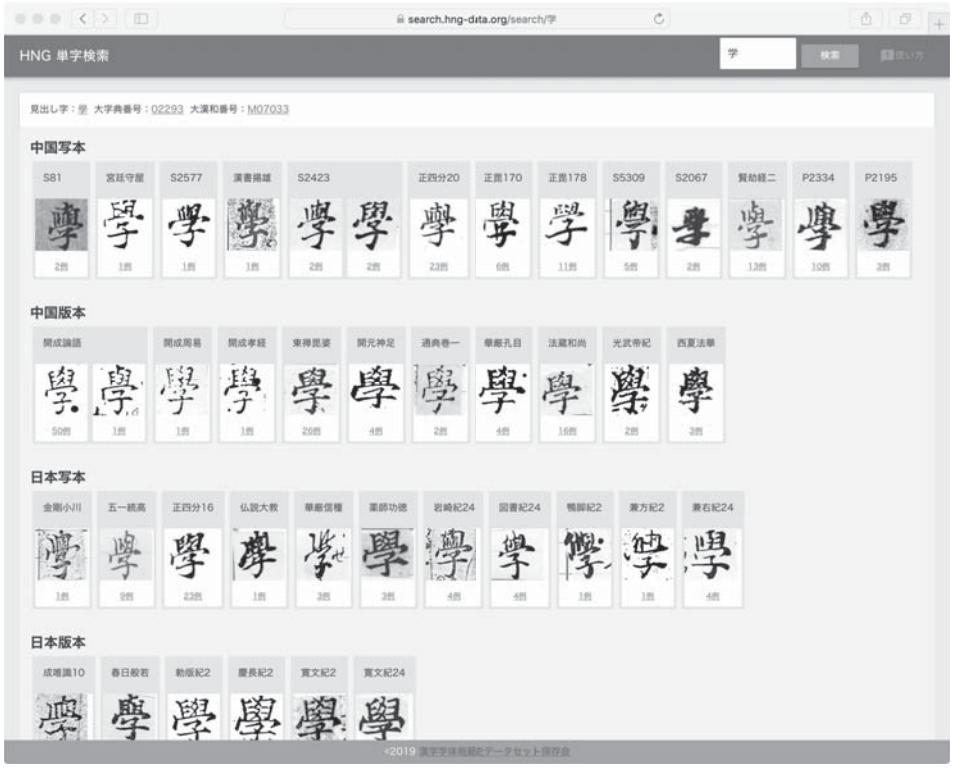

圖 19　HNG 單字檢索の檢索結果の例（學）

6.5.2　檢索結果畫面の構成

檢索結果畫面は大きく分けてヘッダー，見出し字情報，用例情報，フッターからなる。

■ヘッダー部　ヘッダーはトップページと同様なもので，檢索窓から新たな檢索を行うことができる。

■見出し字情報表示部　見出し字情報は檢索した文字に對應する見出し字の情報を簡潔に表示する部分で，「見出し字」，「大字典番號」，「大漢和番號」の値部分はそれぞれ CHISE-wiki における見出し字に對應する UCS 抽象文字オブジェクト，大字典字種（超抽象文字）オブジェクト，大漢和の例示字體オブジェクトのページに對するリンクとなっており，これらのリンクをたどることによってより詳細な情報を見ることもできる。この內，大字典字種オブジェクトは HNG 單字檢索に對應するために大字典データベース［13］［29］をもとに今回新設したものである。

■用例情報表示部　用例情報は HNG における用例を表示する部分で，「中國寫本」，「中國版本」，「日本寫本」，「日本版本」，「韓國資料」，「その他資料」の 6 段に分けて HNG データセットに收錄された各資料での用例が表示される。各資料の項目は上段に略稱，

中央に代表字形，下段に用例數が表示される。また，代表字形の畫像をクリックするとCHISE-wikiにおける代表字形オブジェクトのページに飛ぶことができる。同様に，用例數をクリックするとEsTにおける石塚漢字自體資料の紙カードのページに飛ぶことができる。これらのページを見ることにより例示字形の情報や紙カードの情報を確認することができる。

■フッター部　フッターにはトップページのと同様な檢索窓と漢字字體規範史データセットのWebサイトへのリンクがある。

7 おわりに

データベースの再生のためにはデータ考古學的な手法による『發掘調査』に加え，『データ文獻學』や『データ思想史』とでもいうようなデータを解釋する上でのさまざまな觀點に基づくデータの『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，データセットを繼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）データを十分に解釋・翻譯可能なセマンティクスを付與する必要があると思われる。これは言い替えれば，データベースを永續化するためにはデータセット本體を保存するだけでなく，データを解釋し處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，卽ち，プログラムとして解釋可能なデータという視點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，池田證壽氏，齋木正直氏，劉冠偉氏，北海道大學文學研究科池田證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著者の責任であることはいうまでもない。

參考文獻

[1] Michael Appleby, Tom Crane, Robert Sanderson, Jon Stroop, and Simeon Warner. IIIF Image API 2.1. http://iiif.io/api/image/2.1/, 2016年5月. Version 2.1（Crowned Eagle）.

[2] Juan Benet. IPFS - content addressed, versioned, P2P File System（draft 3）. *arXiv preprint arXiv : 1407.3561*, 2014年.

[3] IIPImage. http://iipimage.sourceforge.net/, 2000-2015年.

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

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

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

[7] John H. Jenkins, Richard Cook, and Ken Lunde (ed.). Unicode han database (unihan), revision 19. http: //www. unicode. org/reports/tr38/tr38-19. html, 2015 年 6 月. Unicode Standard Annex #38.

[8] Tomohiko Morioka. Integration of a chinese character ontology and historical glyph examples. In 9th *International Conference of Digital Archives and Digital Humanities* (*DADH 2018*), pp. 287-300. Taiwanese Association for Digital Humanities / Dharma Drum Institute of Liberal Arts, 2018 年 12 月.

[9] Openseadragon. https: //openseadragon.github.io/.

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

[11] Richard Cyganiak, DavidWood, and Markus Lanthaler (ed.). RDF 1.1 concepts and abstract syntax. https: //www.w3.org/TR/2014/REC-rdf11-concepts-20140225/, 2014 年 2 月. W3C Recommendation 25.

[12] 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, 情報處理學會シンポジウムシリーズ, 第 2007 卷, pp. 237-244. 情報處理學會, 情報處理學會, 2007 年.

[13] 高田智和. 『大字典』データベースをつくる. じんもんこん 2001 論文集, 情報處理學會シンポジウムシリーズ, 第 2001 卷, pp. 221-228. 情報處理學會, 2001 年 12 月.

[14] 石塚晴通, 池田證壽, 岡墻裕剛. 漢字字體規範データベースとその應用. 東洋學へのコンピューター利用 第 17 回研究セミナー, 全國文獻・情報センター人文社會科學學術セミナーシリーズ, 京都大學學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.

[15] 石塚晴通, 高田智和. 漢字字體と文獻の性格との關係 ——「漢字字體規範史データベース (石塚漢字字體資料)」の文獻選定. 石塚晴通監修, 高田智和, 馬場基, 横山詔一 (編), 漢字字體史研究二 —— 字體と漢字情報, pp. 349-359. 勉誠出版, 2016 年 11 月.

[16] 石塚晴通, 高田智和, 守岡知彦. 漢字字體規範史データセット資料一覽. http: //www.hng-data.org/sources.ja.html, 2018 年 7 月.

[17] 石塚晴通, 高田智和, 守岡知彦. 漢字字體規範史データセット保存會. http: //www.hng-data.org/, 2018 年 7 月.

[18] 守岡知彦. Concord：プロトタイプ方式のオブジェクト指向データベースの試み. Linux Conference 抄錄集, Vol. 4, 2006 年.

[19] 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, 情報處理學會シンポジウムシリーズ, 第 2009 卷, pp. 9-14. 情報處理學會, 情報處理學會, 2009 年.

[20] 守岡知彦. CHISE のセマンティック Wiki 化の試み. 情處研報, Vol. 2010-CH-87, No. 8, pp. 1-8, 2010 年 7 月.

[21] 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. じんもんこん 2010 論文集, 情報處理學會シンポジウムシリーズ, 第 2010 卷, pp. 33-40. 情報處理學會, 情報處理學會, 2010 年 12 月.

[22] 守岡知彦. 古典中國語形態素コーパスの linked data 化の試み. じんもんこん 2013 論文集, 情報處理學會シンポジウムシリーズ, 第 2013 卷, pp. 187-194. 情報處理學會, 情報處理學會, 2013 年.

[23] 守岡知彦. CHISE における漢字字體・字形粒度の整理規準について. 東洋學へのコンピューター利用第 26 回研究セミナー, 全國文獻・情報センター人文社會科學學術セミナーシリーズ, pp. 153-190, 2015 年 3 月.

［24］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 ―― 漢字字體規範史データベースの CHISE への收錄を通じて ――．じんもんこん 2015 論文集，情報處理學會シンポジウムシリーズ，第 2015 卷，pp. 1–8. 情報處理學會，情報處理學會，2015 年.

［25］守岡知彦．CHISE による HNG データ收錄の試み．石塚晴通監修，高田智和，馬場基，横山詔一（編），漢字字體史研究二 ―― 字體と漢字情報，pp. 185–203. 勉誠出版，2016 年 11 月.

［26］守岡知彦．データベースの再生と保存についての試論 ―― HNG を例に ――．じんもんこん 2018 論文集，情報處理學會シンポジウムシリーズ，第 2018 卷，pp. 373–380. 情報處理學會，2018 年 11 月.

［27］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．情報處理振興事業協會平成 13 年度成果報告集．情報處理振興事業協會，2002 年. http : //www.ipa.go.jp/NBP/13nendo/reports/explorat/charadb/charadb.pdf.

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

［29］高田智和．大字典データベースをつかう．情處研報，Vol. 2004, No. 58（2004-CH-62），pp. 45–52, 2004 年 5 月.

［30］上田萬年，岡田正之，飯島忠夫，榮田猛猪，飯田傳一（編）．大字典．啓成社，1917 年. http : //kindai.ndl.go.jp/info : ndljp/pid/950498.