様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 ２６ 日現在

機関番号：１４３０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ００６１１

研究課題名（和文）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

研究課題名（英文）Construction of digital archives for historical glyphs of Chinese characters by establishing methods for refining glyph descriptions

研究代表者
守岡　知彦（Morioka, Tomohiko）

京都大学・人文科学研究所・助教

研究者番号：４０３２４７０１

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：石塚漢字字体資料の63資料の情報を漢字字体規範史データセットとして公開した。これは漢字字体規範史データベースが持っていた各資料の字体の情報に加え、その開発過程の情報（文字整理基準の変遷等）もGitリポジトリ化したもので、石塚漢字字体資料の紙カード画像も併せてデータセット化した。また、敦煌文書３資料に対し、フランス国立図書館の電子図書館で公開されているペリオコレクションのIIIF画像を利用して字形の再切り出しを行った。また、従来風のUIを持つ「HNG 単字検索」を開発し公開した。
また、石塚晴通氏他関係者への聞き取り調査を行うとともに、石塚漢字字体資料の調査を行い重要なものに対し保存措置を講じた。

研究成果の学術的意義や社会的意義
漢字字体史研究におけるインフラの一つであったにも関わらず2015年にサービスを停止した漢字字体規範史データベースを今日のネット環境にあった形で再生したことは漢字学・日本語学にとって重要なだけでなく学術情報サービスの再生における貴重なケーススタディーの一つといえる。
また、その内容を自由に利用可能なライセンスをつけたデータセットとして公開したことはその情報を長期保存する上で非常に重要な取り組みといえる。
また、ユーザーコミュニティを立ち上げ、単に技術的な手段だけでなく、社会的な仕組みも含めた保存体制の確立を試みたことも人文系データセットの長期保存において重要な取り組みであるといえる。

研究成果の概要（英文）：We released the information of “Ishizuka Register of Chinese Character Standards of Writing” (IRCCSW; 63 sources version) as the Hanzi Normative Glyphs (HNG) dataset. It is version-controlled using Git and represents information about the HNG database development process (such as glyph classification policy) as well as glyphs and related metadata for each source contained in HNG database. For the images of IRCCSW index cards, we made a corresponding Git repository for each source. For three sources of the Dunhuang manuscripts, the glyph images were re-cut out using the IIIF API for the images of the Perio collection published in Gallica of the National Library of France. We also have developed and released “HNG single character search” which has a UI similar to HNG DB.
In addition, we conducted an interview survey with Professor Emeritus Harumichi Ishizuka and other related parties, and also investigated the storage status of IRCCSW and preserved important materials.

研究分野：漢字情報学

キーワード：漢字字体史　文字情報データベース　データセット保存　データベース再生　画像データベース　IIIF　Linked Open Data　デジタルアーカイブズ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

「漢字字体規範データベース」(HNG) は東アジア漢字圏における文字規範の時代的・地域的変遷と異化の全体像を提示するとともに、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。元々の HNG は収録するソース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た。

このように HNG は豊富な字体用例によって漢字の標準的な字体や規範の変遷を理解する上で有用な情報を提供しており、漢字字体史研究にとって欠かせないツールのひとつであるといえる。また、2015 年に開始した CHISE と HNG の統合の試みにより、従来の HNG における文字単位の検索に加え、CHISE の持つ漢字構造情報データを利用した部品単位の検索を実現し、例えば、「殳」や「酉」といった部品を含む HNG の漢字を検索することで、これらの部品の字体がどのように変遷したかを把握可能にした。一方で、2015 年の春頃に従来の HNG の Web サービスが停止し、2018 年４月時点においても再開の目途が立たないという問題が生じてしまった。このように、HNG はそのデータセットの持つ潜在的な可能性に対し、その機能とサービスの安定性の点で問題を抱えていたといえる。

今後、このような問題を防ぎ、安定的な維持管理と発展を可能にするためには、 Git のような版管理技術の利用とともに、自由ソフトウェアやオープンデータとして公開し、研究者が自由に利用・改編・再配布可能な状態を実現する必要があるといえる。その上で、安定的な一時配布元の体制を組織し、長期にわたる安定的な公開とさまざまな利用を実現することが急務であると考えるに至った。

２．研究の目的

本研究の目的は「漢字字体規範データベース」 (HNG) で構築された前近代の漢字字体用例データベースを継承・発展し、CHISE の文字処理技術や IIIF 等の画像共有技術等を活用した漢字グリフコーパスを実現することである。

近年、版本や写本、拓本等の全文画像や全文テキストのインターネット上での公開の動きが広まってきているが、こうしたテキスト中の漢字を字体に分類し検索可能にしたデータセットは少なく、字体に着目した研究を行う際にはしばしば目による確認に頼る必要があった。この問題を解決するために、前近代の漢字文献の全文画像中の漢字字形（以下、出現字形と呼ぶ）を CHISE の文字処理手法に基づき字体に整理し各字体に漢字構造記述を付与するとともに、各出現字形と元画像（文献）の対応関係を管理することで、全文画像・テキストと漢字情報を統合したデータセットを構築する。

また、こうしたデータセットを長期にわたり安定的に利用できるようにするために、Git を用いてそのリポジトリを版管理するとともに、自由なライセンスによる公開を目指す。

３．研究の方法

a) 過去の HNG データと関連資料の整理

まず、HNG のデータおよびその元となった石塚漢字字体資料の紙カード画像を整理し、その作成過程の情報や製作意図、関連する発表資料やインタビュー記事等をまとめたデジタルアーカイブズとし、Web サイトで公開するとともに、その機械可読データを Git リポジトリ化し、自由なライセンスを付与したオープンデータとして公開する。

現在、従来の HNG データベースはサービスを停止しており、そのデータを参照することができない。しかしながら、過去のバックアップから幾つかの時点のスナップショットを得ることができたので、これらを整理･校訂し、今後の作業のベースとなるデータセットにまとめる。また、過去の経緯や設計意図、各種資料の情報に関しては、石塚漢字字体資料を作成した石塚晴通氏はじめ HNG の開発関係者からの聞き取り調査も行う。そして、こうした情報や過去の発表資料等を併せた HNG に関するデジタルアーカイブズを構築する。

b) Git リポジトリの構築と GitHub の利用の試み

HNG のデータセットおよび関連するデジタルアーカイブズに含まれる機械可読なデータは版管理システムのひとつである Git を用いてリポジトリ化し維持管理を行うとともに、インターネット上での公開を行う。特定の事業者が提供するサービスへの依存を避けるために、公開され

る Git リポジトリのマスターは git.chise.org もしくは新たに設ける公式サイトで提供するが、より多くのユーザーや開発者とのコミュニケーションを計るために、GitHub 等の開発者に人気のある Git ホスティングサービスとの連係も試みる。

c) 現代的な Web 技術の利用

HNG のデータは、現在、Excel ファイルと画像ファイルからなっており、あまり構造化されていない。CHISE では現在文字オントロジーの RDF 化を進めているが、これには CHISE に統合された今西本妙法蓮華經巻五の 645 字体と守屋本妙法蓮華經巻三の 592 字体、開成石經論語の 1329 字体のデータのみが含まれるだけで、その他の大部分は今の所対象外となっている。Linked Open Data の世界との連係を計るには HNG のデータ全体の RDF 化が不可欠であり、残りの部分の RDF 化を進める。また、Web アプリケーションから利用しやすい API の提供を目指す。

d) 全文画像の切出しと整理

漢字字体史研究に寄与する資料の全文画像を文字単位に切りだし、各字形の座標データや対応する UCS のコードポイントもしくは IDS 等の字体記述情報を収集し、字形用例のデータを構築する。これには IIIF を用いて公開されている内外の各種機関が提供している資料を用いる。必要に応じて新たに収集・スキャンしたものも基本的に IIIF による公開を行う。こうして収集された字形用例を石塚漢字字体資料と同様な手法に基づき、字体や字種に分類し、HNG と同様な ID を付与する。

e) 字形用例の整理と CHISE への統合

HNG および新たに全文画像から切り出された字体用例データの CHISE 文字オントロジーとの統合を進める。各字体用例の例示字形を CHISE の多粒度包摂モデルに基づいて字体粒度で分類し、CHISE 文字オントロジーにおける字体オブジェクトのどれかに包摂させるとともに、そうした HNG の例示字形を包摂した字体オブジェクトに対して字体粒度の漢字構造記述を行う。

４．研究成果

本研究課題の成果は、大別すると、

(1) HNG 関連資料の Web 上での公開とデータセットの Git リポジトリの公開
(2) 従来型 UI に似た現代的 Web サービスの実現
(3) 高品質な切り出し字形用例と HNG/CHISE からの全文画像へのリンクの実現
(4) 漢字構造記述に関わる基礎理論の研究及び CHISE における実装

に分けられる。

項目 (1) については、Web サイト

http://www.hng-data.org/

の開設と、自由な Git ホスティングシステムである GitLab を用いた HNG データセット及び関連するデータセットやツールを共同で版管理・公開するための場である

https://gitlab.hng-data.org/HNG/

の開設からなる。

旧 HNG 及び石塚漢字字体資料のうち、公開可能な全 63 資料の Git リポジトリ化を完了し、CC BY-SA 4.0 と GNU GPL version 2 以降の選択制によるオープンデータライセンス・自由ソフトウェアライセンスを付与したオープンデータとして公開を行った。この Git リポジトリ化に際しては、漢字字体規範史データベース（旧 HNG DB）のバックアップデータを整理しその分析結果や関係者からの聞き取り調査の結果に基づき、石塚漢字字体資料に収録されていた 63 資料に対して、代表字形画像とメタデータからなる漢字字体規範史基本データセット（以下、基本データセット）と１資料毎の紙カード画像用リポジトリ（石塚漢字字体資料データセット）に分けて Git リポジトリ化を行った。

基本データセットはその Git リポジトリを

https://gitlab.hng-data.org/HNG/hng-basic-data

で公開するとともに、GitHub 上に

https://github.com/chise/hng-basic-data

というミラーも設けた。基本データセットは、旧 HNG DB のバックアップデータの解析結果に基づき、各時点でのタイムスタンプを保存した形で Git リポジトリ化を行い、各時点のスナップショットを対応するタグやブランチで参照できるようにした。

また、石塚漢字字体資料データセットは

https://gitlab.hng-data.org/HNG/hng-cards_*nn*_*sid*

という形式の Git リポジトリに１資料毎に格納されている。但し、*nn* は HNG のフォルダーコード（http://www.hng-data.org/sources.ja.html の表における（code）欄にある 10 進 2 桁

の数値)で、*sid* は同表の ID 欄にある 3 文字の資料 ID を示す。例えば、誠實論巻八(P.2179)の場合、その URL は

https://gitlab.hng-data.org/HNG/hng-cards_01_jou

となる。

項目 (2) は、「HNG 単字検索」

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

の開発・公開がこれに当たる。これは旧 HNG DB に似た UI を提供するとともに、レスポンシブ Web デザインを実現しており、パソコンのみならずスマートフォンやタブレットなどの多様なデバイス上でもそれぞれのデバイスの画面に適した表示が可能となった。この成果に関しては情報処理学会人文科学とコンピュータ研究会(2019 年 5 月 11 日)と日本語学会 2019 年度秋季大会で HNG データセットに関するものと併せて発表を行った。

また、これに関連して、石塚漢字字体資料および HNG データセットの収録文字の整理に用いられた大字典関連データの Git リポジトリ化も行い、大字典データセット

https://gitlab.hng-data.org/HNG/daijiten-data

として公開するとともに、CHISE との統合を行い、その成果を「じんもんこん 2019」で発表した。

また、文学研究資料館・国立国語研究所・台湾中央研究院歴史語言研究所/数位文化中心とともに「史的文字データベース連携システム」に参加し、歴史的な漢字字形資料の国際的な統合検索の一翼を担うこととなった。

項目 (3) は、HNG の字体データとそのソースである原資料の全文データのリンクを実現するために、試験的にフランス国立図書館の電子図書館 Gallica で公開されているペリオコレクション（敦煌文書）のうち P.2334（妙法蓮華経）と P.2195（妙法蓮華経）の文字を１つづつ切り出したデータ(切り出し字形データ)を HNG データセットと対照可能な形でデータセット化し、CHISE と統合したものであり、これにより「HNG 単字検索」や「CHISE-IDS HNG 漢字検索」の検索結果をたどることで字体データや「石塚漢字字体資料」の紙カード画像と IIIF Image API を用いて表示したペリオコレクションの切り出し字形を比較対象可能な形で表示することができた。この切り出し字形データは

https://gitlab.hng-data.org/HNG/hng-kiridashi-data

で公開した。なお、この切り出し字形データの作成には人文情報学研究所の永崎研宣氏が開発した「切り出しくん」を利用している。氏の協力に感謝したい。また、CHISE との統合結果は CHISE-wiki (EsT) の一部として公開されている。

項目 (4) は、CHISE における字体記述、特に、漢字構造の分析及びその適切な記述法に関する方法論の確立及びその実装に関るものであり、

「内容アドレッシングを用いた多粒度漢字構造情報表現の試み」, 情報処理学会論文誌 61(2) pp.171-178 2020 年 2 月

では多粒度漢字構造情報のより適切な表現法について検討し、これに基づき、 IWDS-1 の包摂規準に基づいた漢字構造情報の正規化システムを試作した。また、

「漢字構造変換の試み」, じんもんこん 2020 論文集, pp.197-202, 2020 年 12 月

及び

“Viewpoints on the Structural Description of Chinese Characters”, Grapholinguistics and Its Applications Vol.5, pp.683-712, 2021 年 2 月

では機能的漢字構造の皮相漢字構造への変換法を提案するとともに、これに基づく自動変換システムを実装し、CHISE 文字オントロジーの構築時にこれを実行し、CHISE 上において両者を複合した検索を実現した。また、後者の論文では漢字構造の機能的部品・漢字構造としての適切さを評価するための指標を提案し、漢字構造及び字体記述の客観的な評価のための基礎を実現したといえる。

５．主な発表論文等

〔雑誌論文〕　計11件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　11件）

１．著者名 Tomohiko Morioka	４．巻 5
２．論文標題 Viewpoints on the Structural Description of Chinese Characters	５．発行年 2021年
３．雑誌名 Grapholinguistics and Its Applications	６．最初と最後の頁 683 - 712
掲載論文のDOI（デジタルオブジェクト識別子） 10.36824/2020-graf-mori	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 守岡 知彦	４．巻 33
２．論文標題 CHISEのWeb API化の試み、ついでに、RDF化四度目の正直?	５．発行年 2021年
３．雑誌名 東洋学へのコンピュータ利用	６．最初と最後の頁 69 - 87
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 守岡 知彦	４．巻 2020
２．論文標題 漢字構造変換の試み	５．発行年 2020年
３．雑誌名 じんもんこん2020論文集	６．最初と最後の頁 197 - 202
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 守岡 知彦，劉 冠偉，高田 智和	４．巻 2019-CH-120 (2)
２．論文標題 漢字字体規範史データセット用従来型UI再生の試み	５．発行年 2019年
３．雑誌名 情報処理学会研究報告 人文科学とコンピュータ (CH)	６．最初と最後の頁 1 - 6
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 守岡 知彦	4．巻 31
2．論文標題 漢字字体の包摂規準の衝突評価の試み	5．発行年 2019年
3．雑誌名 東洋学へのコンピュータ利用	6．最初と最後の頁 51 - 57
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 守岡 知彦	4．巻 2019
2．論文標題 大字典データベースの CHISE との統合の試み	5．発行年 2019年
3．雑誌名 じんもんこん2019論文集	6．最初と最後の頁 51 - 56
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 守岡 知彦	4．巻 94
2．論文標題 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	5．発行年 2019年
3．雑誌名 東方學報	6．最初と最後の頁 320 - 284
掲載論文のDOI（デジタルオブジェクト識別子） 10.14989/250681	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 守岡 知彦	4．巻 61 (2)
2．論文標題 内容アドレッシングを用いた多粒度漢字構造情報表現の試み	5．発行年 2020年
3．雑誌名 情報処理学会論文誌	6．最初と最後の頁 171 - 178
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 守岡 知彦	4．巻 2018
2．論文標題 データベースの再生と保存についての試論―HNG を例に―	5．発行年 2018年
3．雑誌名 じんもんこん2018論文集	6．最初と最後の頁 373-380
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 ―

1．著者名 守岡 知彦	4．巻 30
2．論文標題 漢字構造情報のIPLD化の試み	5．発行年 2019年
3．雑誌名 東洋学へのコンピュータ利用	6．最初と最後の頁 287-301
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 ―

1．著者名 高田 智和	4．巻 2019年1月号
2．論文標題 漢字字体規範史データセット	5．発行年 2019年
3．雑誌名 印刷雑誌	6．最初と最後の頁 78-79
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 ―

〔学会発表〕　計9件（うち招待講演　6件／うち国際学会　2件）

1．発表者名 守岡知彦
2．発表標題 漢字字体規範史データセットにおける版管理
3．学会等名 シンポジウム「字体資料共有の現在と未来」（招待講演）
4．発表年 2021年

１．発表者名
Morioka, Tomohiko

２．発表標題
Viewpoints on the Structural Description of Chinese Characters

３．学会等名
Grapholinguistics in the 21st Century 2020（国際学会）

４．発表年
2020年

１．発表者名
守岡知彦，劉冠偉，高田智和

２．発表標題
漢字字体規範史データセットと単字検索

３．学会等名
日本語学会2019年度秋季大会予稿集

４．発表年
2019年

１．発表者名
高田智和

２．発表標題
漢字字体規範史データセット単字検索

３．学会等名
NINJALセミナー「日本語研究の基盤としての言語資源」（招待講演）

４．発表年
2019年

１．発表者名
守岡知彦

２．発表標題
コンピュータは説文解字の夢を見るか？ －漢字の知識表現と用例－

３．学会等名
2019 年度 富山大学人文学部 中国言語文化講演会（招待講演）

４．発表年
2019年

１．発表者名
高田智和

２．発表標題
中世多摩の文字づかい―板碑と経典文字からわかること―

３．学会等名
多摩郷土誌フェア（招待講演）

４．発表年
2020年

１．発表者名
高田 智和

２．発表標題
『石塚漢字字体資料』と『漢字字体規範史データベース』

３．学会等名
シンポジウム「文字情報データベースの保存と継承」（招待講演）

４．発表年
2018年

１．発表者名
守岡 知彦

２．発表標題
漢字字体規範史データセットの構築・共有計画について

３．学会等名
シンポジウム「文字情報データベースの保存と継承」（招待講演）

４．発表年
2018年

１．発表者名
Morioka, Tomohiko

２．発表標題
Integration of a Chinese character ontology and Historical Glyph Examples

３．学会等名
9th International Conference of Digital Archives and Digital Humanities（国際学会）

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

漢字字体規範史データセット http://www.hng-data.org/ Git repositories of HNG dataset https://gitlab.hng-data.org/HNG HNG単字検索 https://search.hng-data.org/ 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 シンポジウム「文字情報データベースの保存と継承」―漢字字体規範史データセット保存会設立記念イベント― http://www.hng-data.org/events/2018-07-21.ja.html

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	高田　智和 (TAKADA Tomokazu) (90415612)	大学共同利用機関法人人間文化研究機構国立国語研究所・言語変化研究領域・准教授 (62618)	

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関