漢字字体規範史データセットの構築・共有計画について

守岡 知彦

京都大学 人文科学研究所
東アジア人文情報学研究センター

2018 年 7 月 21 日

概要

- 漢字字体規範史データセットとは
- データセットの発掘と整理
- 永続化と利活用
- 将来展望

漢字字体規範史データセットとは

- HNG のコアとなる石塚漢字字体資料をデータセット（オープンデータ）化したもの
- Git リポジトリ (https://gitlab.hng-data.org/HNG/hng-data) で版管理し今月より公開開始
- 石塚漢字字体資料のカード画像と（そこから切り出された）文字画像とメタデータを提供
- 現在の所、48 資料を収録

沿革 (1)

- 2014 年 9 月：CHISE と HNG の連係に関するプロジェクトが始まる (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)
- 2015 年 4 月頃?：HNG がサービス停止
- 2015 年 9 月：HNG と CHISE の統合の試みとして、長安宮廷写経の例示字形データの CHISE 文字オントロジーへの収録作業開始。また、HNG 代替サービスとしてそれ以外の HNG データをとりあえず CHISE-wiki 上で表示するための作業も開始（HNG 48 資料の Git リポジトリ化)。

沿革 (2)

- 2015 年 10 月：HNG 48 資料の CHISE への取込作業が完了。
- 2015 年 11 月：CHISE-IDS HNG 漢字検索を公開。
- 2016 年 4 月：石塚漢字字体資料の紙カード画像の表示機能を CHISE-wiki に追加。
- 2016 年 6 月：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を追加。

沿革 (3)

- 2017 年夏頃：HNG のコア部分の保存計画に関する議論が始まる。
- 2018 年 1 月：HNG データセット保存会構想
- 2018 年 4 月：石塚漢字字体資料の長期保存と発展を目的とした科研費基盤研究 (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」が採択され、データセット保存を目的としたプロジェクト開始

データセットの発掘と整理 (1)

- 一度止まってしまったサービスを復元するのは難しい
  - → 現物が見れない
    - 挙動が判らない
    - どれが正しいデータか判らない
    - データの意味が判らない
  - → 発掘したデータに対する『発掘調査』や『資料批判』『校訂』等の作業が必要になる

データセットの発掘と整理 (2)

- 2014 年 9 月当時は HNG が動いていたので、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だけをもらって（手元に置いて）眺めながらどうするか考えていた
- HNG が止まってから古いバックアップのデータ（2007 年 3 月頃のもの?；48 資料版）をもらって高田さんにいろいろ聞きながらうろ覚えの記憶と想像で頑張ったが良く判らない部分は試行錯誤するしかなかった

データセットの発掘と整理 (3)

- その後、もう少し新しいデータ（2010 年 3 月頃のバックアップ?）を頂いたが中身がややカオス
- 調査してみると過去の作業ミス（番号の取り違え等）や ID や名称等の異同も見つかる

データセットの発掘と整理 (4)

- 48 資料（2006 年度）版をベースにファイル名整理し、Excel ファイル（メタデータ）を CSV 化して Git リポジトリ化 (https://gitlab.hng-data.org/HNG/hng-data)
- 64 資料（2010 年度）版を整理して比較したいが…

永続化と利活用 (1)

- HNG の停止後、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
- Web サービスが止まってもデータがあれば何とかなるはず

  → HNG をデータセット化して、その正式な配布元となる Web サイトを作って公開

  → 組織の改組や研究者の移動等によって URL が変わらないように独自の URL (hng-data.org) を確保

永続化と利活用 (2)

- GitLab の導入
  - Git による版管理
  - 営利企業による占有的なサービスへの依存（例：GitHub）はやめる
  - 利用者が自由に参加・離脱・フォークできるようにする

永続化と利活用 (3)

- 競争的資金の利用と非依存
  - お金と手間がかかる部分は競争的資金で
    - リファクタリング、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
  - 長期間持続可能な体制を目指す

→ メンテナンスコストを下げる

→ データセット保存会

永続化と利活用 (4)

- 非中央化
  - 独自ドメイン (hng-data.org) を取って URL の永続化に努めてもその永続性は人間の努力に依存する部分が生じてしまう
  - 特定サーバーに依存することの問題（非効率、コントロール等の問題）

→ IPFS (InterPlanetary File System) の利用

まとめ：HNG の未来のために

- データを整理し（技術的・ライセンス的に）いじりやすいデータセットを開発・提供することを目指す（とりあえず、https://gitlab.hng-data.org/HNG/hng-data を本日公開）
- HNG の開発過程やそこでの知見、暗黙知等をデータ発掘、物の調査、関係者へのインタビュー等によってアーカイブズ化し、（永続的なデータという意味での）デジタルアーカイブズにすることを目指す
- 全文画像・テキストとのリンクの拡充
- 拡張可能性：新たなデータや技術への対応