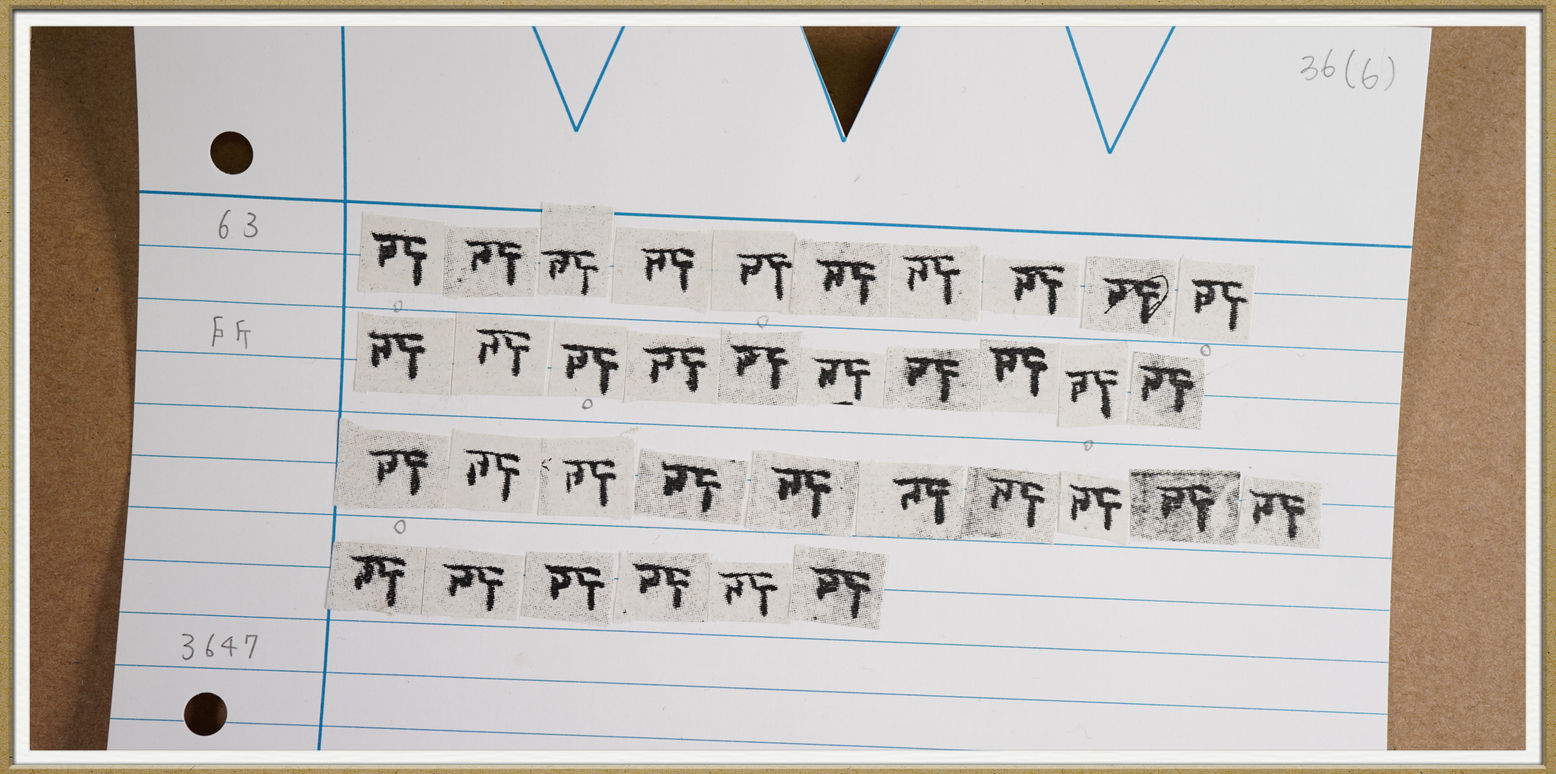


漢字字体規範史データセットにおける版管理

守岡 知彦（京都大学人文科学研究所）

2021年3月20日

はじめに

- 漢字字体規範史データセット (HNG dataset) の Git リポジトリ
  - Git：分散版管理システム（版管理以外の意義も大きいが…）
- HNG のデータの形成過程を追いたい
  - HNG に関わる暗黙知のデータ化、歴史資料としての HNG

Git

- 分散型版管理システム
  - 版管理システム：リポジトリ（プログラムやデータ）の変更履歴を管理
  - 分散型
    - さまざまな場所にリポジトリを置ける
    - 異なる場所、ユーザーの変更を取り込むことができる

    → リポジトリを長期間保全する上で不可欠な性質

HNG の Git リポジトリ

- HNG データセット: https://gitlab.hng-data.org/HNG/hng-data
    - 全資料の代表字形画像と紙カード画像とメタデータを収録
    - HNG-basic-data: https://gitlab.hng-data.org/HNG/hng-basic-data
    - 石塚漢字字体資料リポジトリ群: https://gitlab.hng-data.org/HNG/hng-cards_*nn_sid*
- HNG 切り出し君データ: https://gitlab.hng-data.org/HNG/hng-kiridashi-data
- 大字典データセット: https://gitlab.hng-data.org/HNG/daijiten-data

HNG-data の問題

- HNG-data: https://gitlab.hng-data.org/HNG/hng-data
  - 資料毎に石塚漢字字体資料の紙カード画像と各字体の代表字形（紙カードから切り出したもの）を収録
  - 1 つのリポジトリの HNG に収録された資料の情報を収録
  - 約6.3GBとサイズが大きいため、GitHub にミラーできない

Git リポジトリの分割

- 資料毎に分割
  - GitHub にミラーする資料としない資料を選別しないといけない
- 紙カード画像を分離
  - 全ての資料の代表字形画像とメタデータを GitHub にミラー可能
  - 紙カード画像の GitHub へのミラーは諦める
    - 今回はこの方針を採用
- 紙カード画像本体を IPFS に置く
  - 紙カード画像への参照（ハッシュ値）とメタデータも GitHub に置ける

HNG basic data

- HNG に収録された全ての資料の代表字形画像とメタデータを収録

  https://gitlab.hng-data.org/HNG/hng-basic-data
  (https://github.com/chise/hng-basic-data)

  - カード画像を取り除いで小さくした (6.3GB → 1.6GB)

  - HNG の構築の歴史をタグ・ブランチで表現

    - 現在、HNG16 (2006年2月6日版）〜v.3.9（2008年3月7日版）までの 9 つのバージョンを収録済

石塚漢字字体資料リポジトリ群

- 1つの Git リポジトリ に1つの資料の紙カード画像を収録

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

  nn は資料コード（2桁の数字）

  sid は資料 ID：cf. http://www.hng-data.org/sources.ja.html

  例：今西本妙法蓮華経巻五

  https://gitlab.hng-data.org/HNG/hng-cards_10_khi

- Git LFS (Large File Storage) を利用

HNG 切り出し君データ

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

- IIIF で公開されている全文画像中の字形の座標データ
  - Gallica の全文画像から切り出した字形を表示できる。

  例: https://www.chise.org/est/view/character/repi.hng-myz=5040

- 「切り出し君」に関しては永崎研宣氏の講演をお聞きください

HNG 開発過程のリポジトリ化

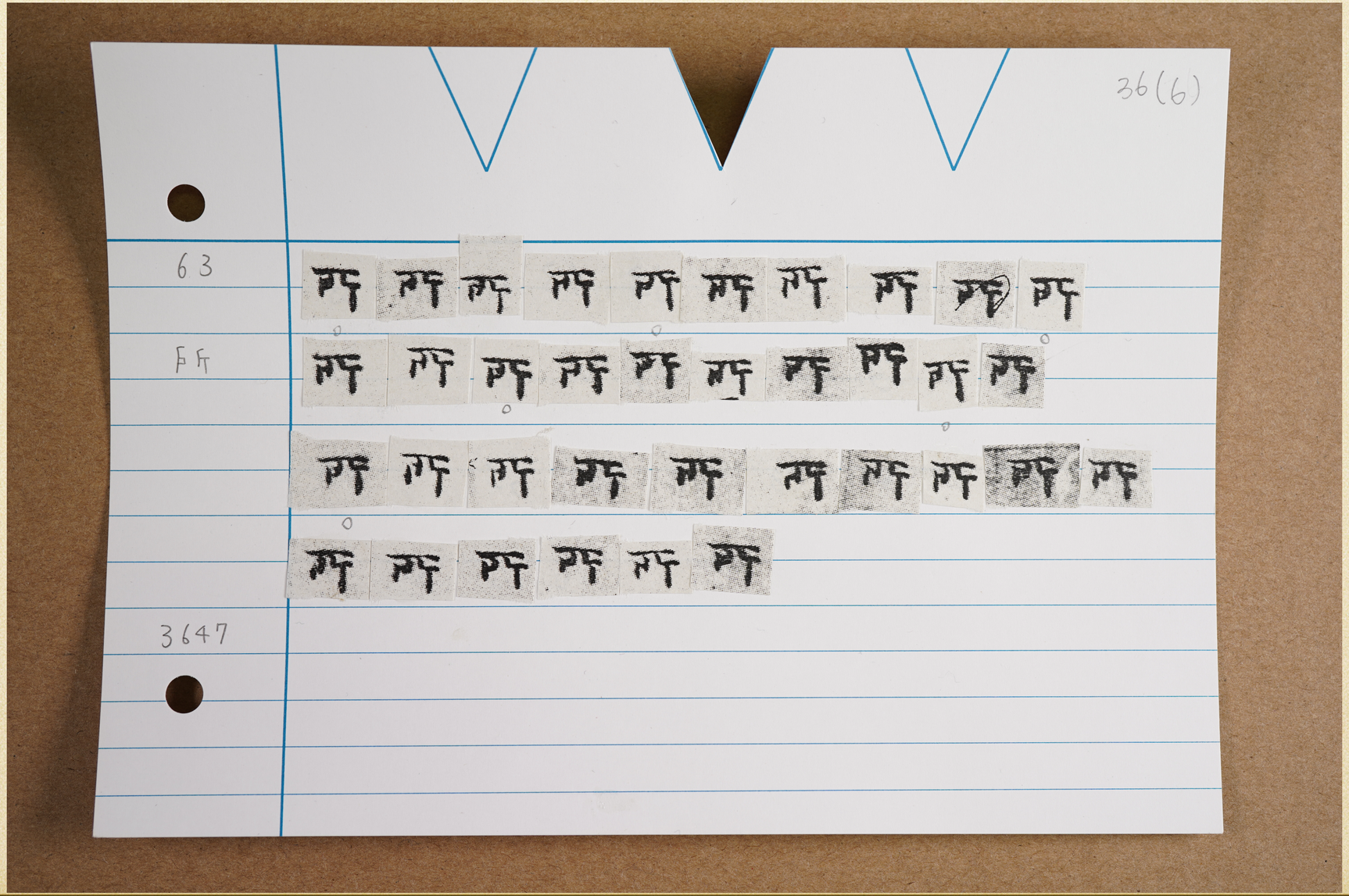
- HNG には独自の文字整理基準・字体判定基準があるがその情報は機械可読化されておらずまた揺れもある。また資料選定（優先度）の情報も重要
  - これらのいくつかは開発中に明確化されたり変更されたりした
  - 当事者にとっての常識は意識されづらい（暗黙知）
- HNG は十数年（紙カード時代も含めるともっと長い）かけて開発された長いプロジェクト
  - もし HNG の開発に Git を使ってたら？というイメージで再構築
    - タイムスタンプ を利用。作業者は不明なので HNG Developers ということにした。

さまざまなデータと差異

- タイムスタンプ とメタデータの差異は重要なヒントになるが、メモや補助的なデータ・ツールの類もヒントになる
- 紙カードの画像と紙カードの原本の差
- 紙カード上の書き換えの痕跡

https://www.chise.org/est/view/image-resource/rep.id=HNG...012...card...0620

と



まとめ

- HNG データセットの Git リポジトリを再構築中
  - 代表字形とメタデータからなる基本データリポジトリ https://gitlab.hng-data.org/HNG/hng-basic-data と 1 資料毎の紙カード画像用リポジトリ https://gitlab.hng-data.org/HNG/hng-cards_*nn_sid* に分割
  - HNG の開発過程の情報の表現
- IIIF で公開された全文画像とのリンクの拡充
  - 従来の京大人文研所蔵開成石経拓本に加え、「切り出し君」のデータの取り込み

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