CHISE における HDIC サポートの現状と課題

守岡 知彦　（京都大学人文科学研究所 / CHISE project）

2023-01-21 (HNGデータセット保存会第４回シンポジウム「古辞書データ共有と拡張」)

はじめに

- CHISE に平安時代漢字字書総合データベース (HDIC) の古字書データをくっつける
  - これができると CHISE 経由で
    HDIC と漢字字体規範史データセット (HNG) もくっつく

CHISE　CHaracter Information Service Environment

- 漢字を中心とした文字オントロジーと関連する Web サービスなどを提供
  https://www.chise.org/
  - https://gitlab.chise.org/CHISE （Git リポジトリ）
  - https://www.chise.org/ids-find （CHISE IDS 漢字検索）
  - https://www.chise.org/est/view/character/字 （E_ST (CHISE-wiki) 頁の例）
    - http://api.chise.org/character/v1/get-info?character=字 (char-info API の例)
    - https://api.chise.org/v0/character/ids-match?ids=□亠□xx (IDS-match API の例)

平安時代漢字字書総合データベース (HDIC)

漢字字体史研究と漢字字書編纂史研究とに資することを目的に池田証壽氏（北海道大学名誉教授）らが開発している中国・日本の漢字字書（古字書）のデータベース https://hdic.jp/
https://github.com/shikeda/HDIC (データセット)
https://viewer.hdic.jp/ (HDIC Viewer by 劉冠偉氏)

- 中国
  - 『玉篇』(梁・顧野王撰，543 年成，現存 2,087 字)(原本玉篇;YYP: Yuanben Yupian database)
  - 『大広益会玉篇』(宋・陳彭年等撰，1013 年成，約 22,800 字)(宋本玉篇;SYP: Songben Yupian database)
  - 『龍龕手鏡』(遼・行均撰，997 年成，約 26,000 字，高麗版;GLS: Gaoliben Longkan Shoujing database)
- 日本
  - 『篆隷万象名義』(空海撰，9世紀初，約 16,000 字) (KTB: Kosanjibon Tenrei Bansho Meigi database)
  - 『新撰字鏡』(天治本，昌住撰，10 世紀初，約 21,000 字)(TSJ: Tenjibon Shinsen Jikyo database)
  - 『類聚名義抄』(図書寮本，11 世紀初，約 3,600 項目) (ZRM: Zushoryobon Ruiju Myogisho database)
  - 『類聚名義抄』(観智院本，12 世紀後半，約 32,000 項 目)(KRM: Kanchi’inbon Ruiju Myogisho database)

HDICの目標

漢字字体史研究と漢字字書編纂史研究とに資すること

- 字体史研究のために
  - 掲出字画像データベース
    - HNG との連携
- 字書編纂史研究のために
  - 全文テキストデータベース
    - 字書（引用箇所）の比較

当初考えられていたユーザーインターフェースのイメージ

古字書掲出字 の CHISE への統合 (1)

- HNG と同様なグリフコーパスとしての側面に着目
- HDIC の掲出字を、HNG の代表字形と同様に、字形粒度のオブジェクトとして CHISE 文字オントロジーに取り込む
  - 但し、対応する字体粒度のオブジェクトも定義可能にする
    - 多粒度漢字構造モデルに基づききちんと包摂階層に位置付ける場合
    - 対応する UCS 抽象文字がまだ存在してない場合
  - HDIC の掲出字に対応する字形粒度のオブジェクトと UCS 抽象文字（なかったら対応する字体粒度オブジェクト）との間に関係素性を張る

古字書掲出字 の CHISE への統合 (2)

- E_ST 上で HNG の例示字形と HDIC の掲出字が 並んで表示される
  - （HNG には存在しない利用頻度は低いけど伝統的な漢字字書によくある字（の初唐標準字形）が補完されてうれしい気がする）
- UCS 抽象文字が存在せず IDS で記述された文字も CHISE IDS 漢字検索で検索できる

古字書掲出字 の CHISE への統合 (3)
篆隷万象名義の篆書掲出字の統合

- 篆隷万象名義の篆書掲出字は現代に伝わる説文小篆と結構違う
  - 説文解字の唐代の写本の断簡にある懸針体と同様の特徴を持っている

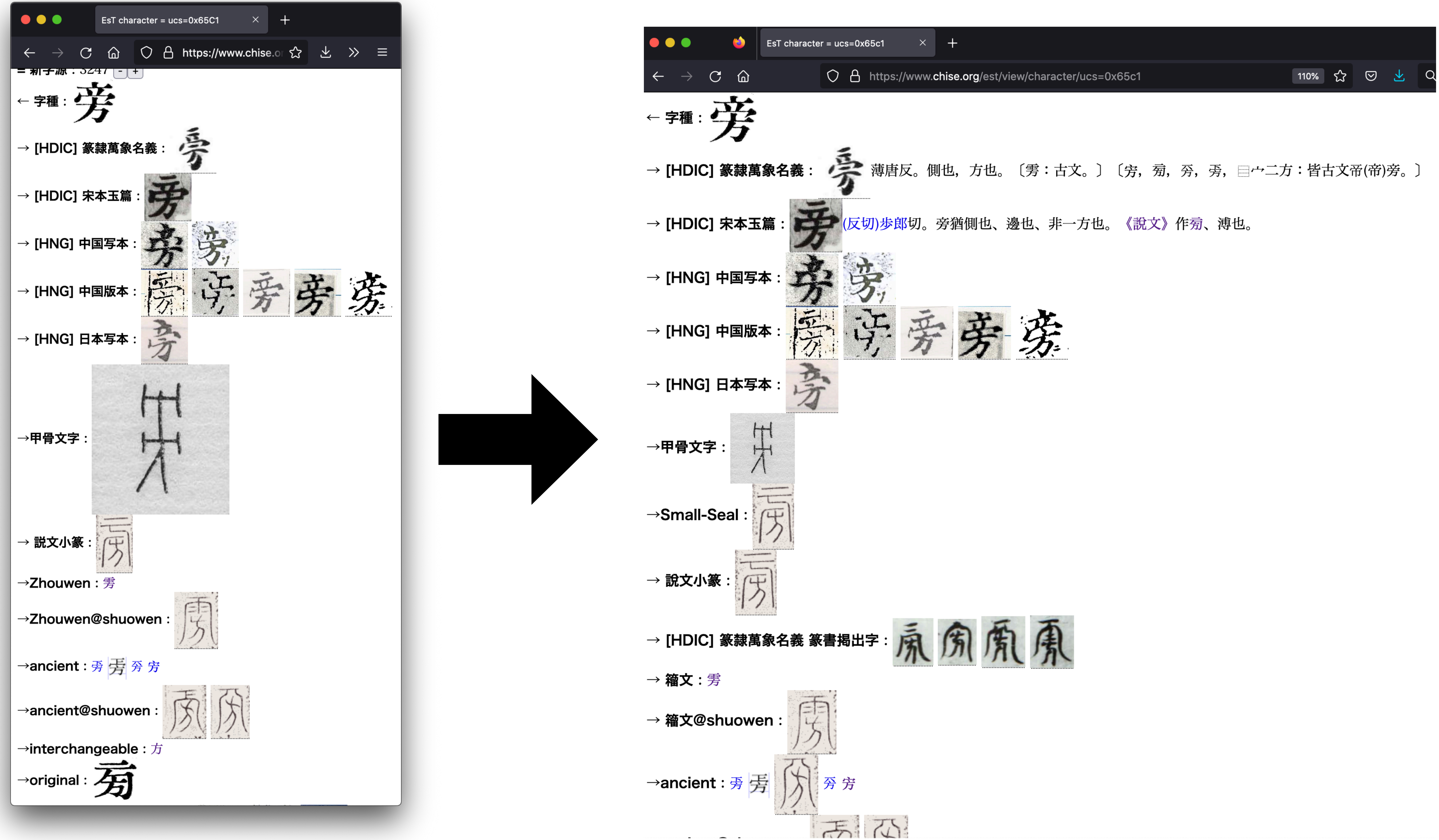
  ➡開成石経規範字体に影響を与えた唐代の説文を知るヒントのひとつ

- 既存の汲古閣本系の説文小篆と大きく異なる場合は両者を包摂する抽象文字オブジェクトを新設。差異が軽微な場合は同じ字体オブジェクトに包摂される字形オブジェクトとした（素性名：===chise-hdic-ktb-seal ）

- 李媛氏から頂いた高山寺本篆隷万象名義の篆書掲出字のデータを利用
  - 低解像度だったため、国会図書館の画像を利用して OpenCL の物体認識で切り出し
    - データ：https://github.com/chise/HDIC-heading-character-coordinates

古字書注文の CHISE への統合 (1)

- 少し前から CHISE に大字典データベース（by 高田智和氏）と汲古閣本系説文解字説文解字（京大人文研所蔵の２つの版本を鈴木俊哉氏が切り出したデータ）を入れた結果、CHISE が字書的にも使えるようになった（気がする）
- HDIC には注文のデータがあり、一部は全文画像へのリンクもあるので、せっかくならこれらも使いたい
  - HDIC の辞書項目に対応するオブジェクトに HDIC のデータを素性値として突っ込めば良い
    - 今回は HDIC の掲出字字形オブジェクトで代用



EsT character = ucs=0x65C1
https://www.chise.o
← 字種 :
→ [HDIC] 篆隸萬象名義 :
→ [HDIC] 宋本玉篇 :
→ [HNG] 中国写本 :
→ [HNG] 中国版本 :
→ [HNG] 日本写本 :
→甲骨文字 :
→ 說文小篆 :
→Zhouwen : 雱
→Zhouwen@shuowen :
→ancient :
→ancient@shuowen :
→interchangeable : 方
→original :
EsT character = ucs=0x65c1
https://www.chise.org/est/view/character/ucs=0x65c1
110%
← 字種 :
→ [HDIC] 篆隸萬象名義 : 薄唐反。側也，方也。〔雱：古文。〕〔㝑，旁，㫄，雱，▤宀二方：皆古文帝(帝)旁。〕
→ [HDIC] 宋本玉篇 : (反切)步郎切。旁猶側也、邊也、非一方也。《說文》作旁、溥也。
→ [HNG] 中国写本 :
→ [HNG] 中国版本 :
→ [HNG] 日本写本 :
→甲骨文字 :
→Small-Seal :
→ 說文小篆 :
→ [HDIC] 篆隸萬象名義 篆書掲出字 :
→ 籀文 : 雱
→ 籀文@shuowen :
→ancient :

古字書注文内容の CHISE への統合

- 古字書の注文には漢字の形音義等の解説が含まれ、CHISE 文字オントロジーの記述内容と重なる部分が存在する（例えば、異体字関係の情報やその典拠情報）
  - 但し、古字書の注文は自然言語（古典中国語や昔の日本語等）で書かれているのに対し、CHISE はラベル付き有向グラフで書かれている
    - ➡ 古字書の注文を解析し、その意味内容を CHISE の形式に翻訳する必要がある
      - 宋本玉篇(SYP)の注文データはぱっと見綺麗なので機械処理して異体字関係のデータも取り出そう

古字書注文の自動解析

- 今回はルールベースで
  - MeCab 漢文も UD 漢文も（多分、現存する機械学習ベースの古典中国語用自然言語処理は概ね全部）字や音に対する注釈が超苦手（漢字を見たら字そのものじゃなくてその字が表す語（や意味）として解釈する。反切（漢字２文字で音節を示す表記）やその他音注も同様。たまたま名詞に誤解したら良いけどそうじゃないとぼろぼろに崩壊しがち）
    - でも例外も少なくない（たかだか有限個だから例外を見つけるたびにルールをいじれば原理的には解決できるはず）

難しい例

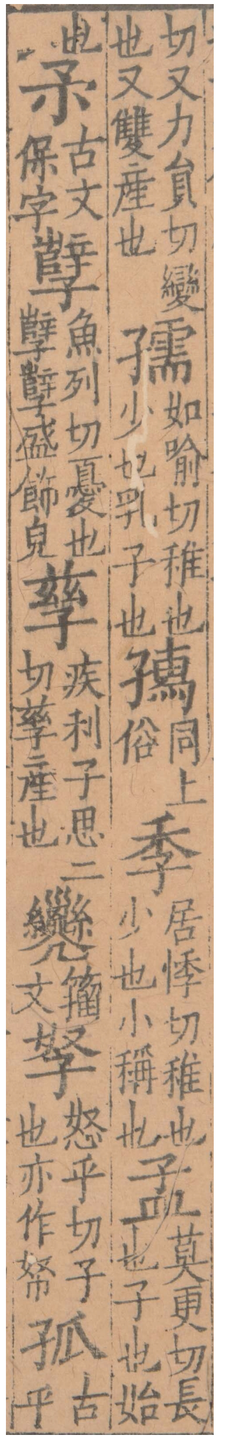
- 「孟」の項目の次に「𡥈」の項目があり
  その注文に「古文保字。」とある。
    - 他の例と同様に考えれば

$$𡥈 \xleftarrow[\text{古字}]{} 保$$

なのだがこのケースの場合

$$孟 \xrightarrow[\text{古字}]{} 𡥈 \xleftarrow[\text{古字}]{} 保$$

（多分、説文で𡥈(保)と𠫓(孟)が重複してることを暗示している?）



おわりに

- CHISE に HDIC のうち、天治本新撰字鏡 (TSJ), 宋本玉篇 (SYP), 高山寺本篆隷萬象名義 (KTB) を統合した
  - HNG と並べて表示できるようになった
  - HDIC Viewer へのリンクを実現した
  - 注文も取り込んだ
    - E_ST の拡張＋presentation-format の設定により、抽象文字のページに HDIC の各字書の注文を並べて表示できるようになった
      （結果的に池田先生が当初考えてたのに近い感じになった？）