

東アジアの歴史的文字関連資料共有のための データインフラストラクチャの構築に関する準備研究

<https://chise.nijl.ac.jp/> (GitLab: <https://gitlab.nijl.ac.jp/CHISE>)

メンバー：守岡知彦（研究代表／国文学研究資料館） ・ 劉冠偉（京都大学人文科学研究所）

1. 研究の目的

本研究の目的は、漢字研究にとって重要な情報資源である

- ・漢字字体規範史データセット (HNG)
<https://www.hng-data.org/>
- ・平安時代漢字字書総合データベース (HDIC)
<https://hdic.jp/>
- ・文字情報サービス環境 CHISE
<https://www.chise.org/>
- ・国書データベース（日本古典籍くずし字データセット）

をはじめとする漢字関連データセットの長期維持のために、複数の地域に分散された分権的な情報基盤の実現を目指すものである。

2. 背景

人文学の研究成果は伝統的には論文や本の形で刊行されてきた。一方、デジタル人文学や人文情報学を採り入れた新しい形としてデータベース、特に、近年は Web サービスとしての公開が進んでいる。しかしながら、前者に比べて後者の長期保存の難しさが徐々に認識されるようになってきた。

3. サブプロジェクト

漢字系データセットの連携手法の研究

漢字を中心とした東アジアの文字関連資料に関するデータセットを互いに連携する手法を検討。

→歴史的漢字字体や古辞書やその他古典籍に関するデータセットを統合し漢字の機械処理の精度向上や適用範囲拡大を目指す



さまざまな字体・字形の〈經〉

日本の古典籍の 自然言語処理に 関する研究

漢字は古典中国語の 1 形態素を 1 文字で表現するために生まれた文字体系。従来、異体字という漢字間の関係とみなされてきた現象の中には形態素の異表記と考えたほうが良いものが少なくない。そこで、漢字の字種と形態素や WordNet の synset との対応関係や、漢文や古典日本語における形態素や構文構造コーパスの構成法・データ化を検討。

『日本古典籍くずし字データセット』
(国文学研究資料館蔵『雨月物語』／CODH加工)



日本語の古典籍における「經」

中国での字体規範の変化や中国語の音韻変化を吸収し日本語を表現するために生まれた多彩な表現やかな文字などの理解

データセットの分権化に関する研究

データ提供組織の改変等で URL が変わるとアクセスできなくなる現行の Web 技術の問題の抜本的解決を目指す。IPFS 等の Web3 技術の人文系学術データへの適用可能性の検討や実証実験



・IPFS 版 CHISE IDS 漢字検索

- ・ <https://ipfs.nijl.ac.jp/ipns/ids-find.chise.org/index.ja.html>
- ・ <https://ipfs.nijl.ac.jp/ipfs/Qme94UnoCa5fXGUNFnVByyf9sWdbSpbUQDhVsJHLvds4DA/index.ja.html>

CHISE の再実装

- ・CL-Concord (*1) & CL-CHISE (*2)
(*1) <https://gitlab.nijl.ac.jp/CHISE/cl-concord>
(*2) <https://gitlab.nijl.ac.jp/CHISE/cl-chise>
- Common Lisp / Valkey (Redis) ベースの新実装。
- 現代的な Web 技術に基づく新たな UI の開発を検討。

InterPlanetary Linked Data (IPLD) ベースのオブジェクト ID の実装
次世代技術のテストベッド