大字典データベースの CHISE との統合の試み

守岡 知彦 （京都大学 人文科学研究所）

多くの漢字字書、漢和辞典の多くは漢文（古典中国語）のテキストを読むための辞書としての性格が強いのに対し、「大字典」は和訓や国字、日本での通用字体を多く採録した日本語のための漢字辞書としての性格を持ち、日本語学（国語学）の分野で重用されてきた。「石塚漢字字体資料」における漢字の整理においても「大字典」が用いられたことからその後身である「漢字字体規範史データセット」のデータやモデルを理解する上でも大字典の情報は重要である。そこで、大字典の情報を特にその字種情報に着目して CHISE 文字オントロジーとの統合を試みた。

An attempt to integrate Daijiten database and CHISE

MORIOKA, Tomohiko (Institute for Research in Humanities, Kyoto University)

While many Chinese character dictionaries and Chinese-Japanese dictionaries are basically designed as dictionaries to understand classical Chinese texts, “Daijiten” has a character as a Chinese character dictionary for the Japanese language, which contains a lot of *wa-kun* (native Japanese readings of characters based on meanings), *kokuji* (Japanese local Chinese characters), and Chinese character glyphs commonly used in Japan. Therefore, it has been heavily used in the field of linguistics for Japanese. Since “Daijiten” was also used to organize the characters in “Ishizuka Register of Chinese Character Standards of Writing”, the information of “Daijiten” is important to understand the data and models of “HNG (Hanzi normative glyphs) dataset” which is its successor. Therefore, we tried to integrate the information of Daijiten with the CHISE character ontology especially focusing on the information of *jishu* (Chinese character category corresponding to conceptual character that include one or more different glyph-forms with the same phonetic value and meaning).

1 はじめに

説文解字や康熙字典のような伝統的な漢字字書、あるいは、近現代に刊行された大漢和辞典や各種漢和辞典のような漢和辞典の多くは漢文（古典中国語）のテキストを読むための辞書としての性格が強く、その多くが（古典）中国学者によって編纂されてきたといえる。それに対し、上田万年、岡田正之、飯島忠夫、栄田猛猪、飯田伝一編「大字典」は国語学者と漢学者の協同作業によって編纂され、国語辞書や前近代の日本の漢字字書とのギャップを埋めることを目指し、和訓や国字、日本での通用字体を多く採録したという点で特異な性格を持っているといえる。現在は絶版であるものの、1917（大正６）年の初版以来、改定重版を重ね普及しており、特に国語学の分野で重用されてきた。また、既に著作権が切れていることから自由なデータを作る基盤としても有用であるといえ、初版と 1920（大正９）年刊行の縮刷版が国会図書館デジタルコレクションに収録されており、Web 上で簡単に閲覧が可能である。

現在、我々は、漢字字体規範史データベース (Hanzi Normative Glyphs; HNG) [4] のデータセット化 (漢字字体規範史データセット) や検索サービスの再構築、新たな技術を用いた試み等を行っているが、漢字字体規範史データベースやその前身となる「石塚漢字字体資料」は大字典によって文字の整理を行っており、そのデータやモデルを理解する上で大字典の情報は重要であるといえる。

大字典のデータベース化の試みとしては高田智和氏による掲出字を対象にしたもの [3][7] があり、劉冠偉氏によるこの和訓拡張の試み [2] がある。ここでは前者をベースに特にその字種情報（異体字シソーラス）に着目して CHISE 文字オントロジー [1] との統合を試みた。

2　データセット化

2010 年 3 月 4 日付の漢字字体規範史データベースのバックアップデータ[*1]に含まれていた大字典データベースの Excel ファイル `daijiten_DB_20061008.xls`[*2*3]をタブ区切り (tab-separated values; TSV) 形式に変換したものをベースに作業を行った。データセット化にあたり Git を用いて版管理を行い、漢字字体規範史データセットと同様に GitLab を用いて `https://gitlab.hng-data.org/HNG/daijiten-data` という Git リポジトリを設け、これをマスターリポジトリーとした。

この Git リポジトリー（以下では、「大字典データセット」と呼ぶことにする）には、現在の所、

- `daijiten_DB.txt`
- `daijiten_page_number.csv`

という 2 つのデータファイルが存在する。

2.1　文字データ

`daijiten_DB.txt` は前述の Excel ファイルから変換された TSV 形式のファイルで、

1. 字種コード
2. 大字典番号
3. 部首番号：大字典の部首番号
4. 文字
5. x0208（区点番号）
6. 包摂・備考
7. UCS
8. 大漢和番号

という欄からなる。これは [7] で解説されているデータ項目と同様であると考えられる。「文字」はおそらくここでのデータ項目『JIS 漢字』(JIS 漢字の包摂字体）相当だと思われる。「包摂・備考」には『文字包摂に関する情報』(JIS 包摂規準適用による連番包摂、互換包摂等を記したもの）と『備考』を 1 つの欄にまとめたものと思われる。また、UCS は JIS X 0221:2001 のものであり BMP の範囲内のものだけが記されている。

2.2　ページ番号データ

`daijiten_page_number.csv` は CSV 形式のファイルで、

1. ページ：大字典本文最初のページを 1 とした時のページ番号
2. 印字頁：国会図書館デジタルコレクション

[*1] 北海道大学で保存されていたもので媒体は外付ハードディスクであった。2015 年 11 月 21〜22 日に開催されたシンポジウム 「字体と漢字情報 —HNG 公開 10 周年記念—」の際に当時著者が持っていた 48 資料版 HNG よりも後に作成されたより完全なバックアップデータがないか関係者に相談した所、北海道大学で保存されていることが判り、石塚晴通氏から提供して頂いたものである。

[*2] ファイル名から推測するに 2006 年 10 月 8 日の版だと思われるがタイムスタンプは 2004 年 4 月 29 日になっている。

[*3] このバックアップデータが収録されていたハードディスクには「石塚漢字字体資料」由来の 64 資料やその選定前の候補となった資料のリスト、作業用のツールやデータに関ると思われるもの、作業途中のデータ等が含まれており、この大字典データベースの Excel ファイルは HNG 作成のための作業用データの一つとして用いられていたものだと思われる。

で公開されている大字典初版、及び、1920年刊行の縮刷版に見えるページ番号。両者に差異がある場合は初版/縮刷版 のように / で区切って並記する。また、不鮮明で読めない場合は ? を記し、ページが欠落している場合は - を記す。

3. 大字典番号：大字典の文字番号

という欄からなる。これは劉冠偉氏が入力したデータ[*4]をベースに、自由な OCR ソフトウェアである Tesseract (Version 4.1.0) を用いて国会図書館デジタルコレクションの版面画像を処理し、ページ番号が極端に飛んだり前後の番号と矛盾するようなものを除外するなどして矛盾がなく確からしい結果だけを採用してデータを補い、また、その一部を手で修正したものである。現在の所、全 2603 ページ中、1577 ページ分の情報が入っている。国会図書館デジタルコレクションで公開されている版面に不鮮明な部分が少なくなく、今後はより鮮明な版面データを入手して作業することを計画している。

3 CHISE との統合

3.1 大字典番号の取り込み

大字典番号に対応する CHISE の ID 素性を表 1 に示す。

包摂粒度	ID 素性名	URL 表現
超抽象文字	==>daijiten	a2.daijiten
字体	=daijiten	rep.daijiten
抽象字形	==daijiten	g2.daijiten
字形	===daijiten	repi.daijiten

表 1　大字典番号に対応する ID 素性

==>daijiten は字種コードに対応するもので、両者の組は大字典データベース（及び、HNG）における字種層の超抽象文字（以下、字種オブジェクトと呼ぶ）を示す。任意の字種オブジェクト C は daijiten_DB.txt での記述に基づき、1 つ以上の大字典字体オブジェクト $G_1, G_2, ...$ に対して、

$$C \texttt{ ->denotational@usage } G_1, G_2, ...$$

という包摂関係を持つ。[*5]

例えば、「丘」に対応する字種オブジェクト

〈*丘*〉 = ((==>daijiten . 26)) [*6]

は字体オブジェクト

「垈」 = ((=daijiten . 1702)) [*7]
「丘」 = ((=daijiten . 26)) [*8]
「北」 = ((=daijiten . 33)) [*9]

との間に

- 〈*丘*〉 ->denotational@usage 「垈」
- 〈*丘*〉 ->denotational@usage 「丘」
- 〈*丘*〉 ->denotational@usage 「北」

という包摂関係を持つ。

なお、ある大字典字体オブジェクトが複数の字種オブジェクトに対応する場合[*10]には

$$G \texttt{ <-denotational@usage } C_1, C_2, ...$$

[*4] 1〜243 頁までを収録している。

[*5] 関係素性 ->denotational@usage は包摂関係を示す素性 ->denotational に慣用的用法を示すドメイン usage を付けたものである。これは通常の包摂関係が基本的に包摂規準に基づくグリフ（抽象形状）に着目したものであり、隷変の仕方や楷書における筆遣い等の差異に基づくものに対し、字種と字体の関係はそれとは異なるメカニズムであることが多いことを鑑みて、別ドメインを付与し区別したものである。

[*6] これは S 式表現であり、対応する URL は http://www.chise.org/est/view/character/a2.daijiten=26 である。

[*7] http://www.chise.org/est/view/character/rep.daijiten=1702

[*8] http://www.chise.org/est/view/character/rep.daijiten=26

[*9] http://www.chise.org/est/view/character/rep.daijiten=33

[*10] 別字衝突していたり、全く別の字義を示すようになり新たな異体字関係が生じていたりするケース等

という記述を該当する字体オブジェクト G に対して行う。

例えば、字体オブジェクト

「攵」 = ((=daijiten . 4223)) [*11]

は字種オブジェクト

〈*支*〉 = ((==>daijiten . 4222)) [*12]
〈*文*〉 = ((==>daijiten . 4321)) [*13]

との間に

- 〈*支*〉 ->denotational@usage 「攵」
- 〈*文*〉 ->denotational@usage 「攵」

という包摂関係を持つ。

3.2 全文画像へのリンク

国立国会図書館デジタルコレクションは啓成社から出版された大字典のうち既に著作権が切れたものを 2 種類収録しインターネット上で公開している。そこでこれらの画像を用いて大字典字体・字種オブジェクトの EgT[5] (CHISE-wiki) ページから国立国会図書館デジタルコレクションで公開されている大字典の全文画像にリンクを張ることにした。

このために、2.2 節で述べた daijiten_page_number.csv のデータに基づき、大字典字体オブジェクト、及び、字種オブジェクトに対し、文字素性 daijiten-pages を付与した。これは該当するオブジェクトに対応する本文ページ番号のリストを値として持つ。

前述のように国立国会図書館デジタルコレクションには 2 種類の大字典が収録されているが、これは

- 大正 6 年刊行の初版
  https://doi.org/10.11501/950498
  (以下では ndl-950498 と呼ぶ)
- 大正 9 年刊行の縮刷版
  https://doi.org/10.11501/950499
  (以下では ndl-950499 と呼ぶ)

である。いずれも IIIF マニフェストが提供されており、その URL はそれぞれ

- https://www.dl.ndl.go.jp/api/iiif/950498/manifest.json
- https://www.dl.ndl.go.jp/api/iiif/950499/manifest.json

である。

両者とも欠落や重複[*14]等によるページの乱れがあるが、調査した所、2.2 節で述べたページ番号を p, ndl-950498 のページ番号を P_{950498}, ndl-950499 のページ番号を P_{950499} とした時、

$$
\begin{aligned}
P_{950498} &= p/2 + 23 \ if \ p < 229 \\
&= p/2 + 24 \ if \ p < 261 \\
&= p/2 + 25 \ if \ p < 263 \\
&= p/2 + 26 \ if \ p < 516 \\
&= p/2 + 27 \ \text{(上記以外)} \\
P_{950499} &= p/2 + 20 \ if \ p < 1317 \\
&= p/2 + 21 \ if \ p < 1325 \\
&= p/2 + 22 \ if \ p < 1327 \\
&= p/2 + 23 \ if \ p < 1366 \\
&= p/2 + 24 \ if \ p < 1482 \\
&= p/2 + 25 \ if \ p < 1932 \\
&= p/2 + 26 \ if \ p < 2225 \\
&= p/2 + 27 \ if \ p < 2241 \\
&= p/2 + 28 \ if \ p < 2257 \\
&= p/2 + 29 \ if \ p < 2451 \\
&= p/2 + 28 \ if \ p <= 2452 \\
&= p/2 + 27 \ if \ p <= 2454
\end{aligned}
$$

[*11] http://www.chise.org/est/view/character/rep.daijiten=4223
[*12] http://www.chise.org/est/view/character/a2.daijiten=4222
[*13] http://www.chise.org/est/view/character/a2.daijiten=4321
[*14] ndl-950498 では 284 と 285 が重複しており、ndl-950499 では 705 と 706, 764 と 765, 991 と 993 が重複している

$$= p/2 + 26 \ if \ p <= 2456$$
$$= p/2 + 25 \ if \ p <= 2458$$
$$= p/2 + 24 \ if \ p <= 2460$$
$$= p/2 + 23 \ if \ p <= 2462$$
$$= p/2 + 22 \ \text{（上記以外）}$$

という関係が成り立つことが判った。但し、$p/2$ は p を 2 で割った商で小数点以下は切り捨てて整数としたものとする。

ndl-950498 の方が落丁が少ないため、前述の p から P_{950498} への関数を用いて、EsT における表示用メソッド

`space-separated-daijiten-page-list`

を定義した。これは対象となる文字オブジェクトの文字素性値を p として P_{950498} を計算し、IIIF ビューアーの TIFY (`https://github.com/subugoe/tify`)[*15]を利用して ndl-950498 の P_{950498} のページを表示するための URL へのリンクを作成するものである。この表示用メソッドを文字素性 `daijiten-pages` の素性属性 `value-presentation-format` に設定することにより、大字典字体・字種オブジェクトの EsT ページから国会図書館デジタルコレクションの本文画像に飛ぶことができた。

4 おわりに

大字典データベースのデータセット化とその CHISE との統合について概説した。

漢字字体規範史データセットを理解する上でその文字・字体の整理に用いられた大字典は重要な資料の一つといえ、これをデータセット化しその Git リポジトリーを提供することは漢字字体規範史データセットを利用する上でも有用であるといえ、また、大字典の日本語学（国語学）分野での重要性を鑑みればそれ単独でも有用な資源であるといえる。た、

大字典データセットを CHISE 文字オントロジーと統合することにより、CHISE の Web サービスや HNG 単字検索 [6] から大字典データセットの情報を利用できるようにした。また、CHISE-wiki 上で大字典の文字のに対応するページから全文画像へのリンクを実現することができた。現在の所、全体の約 60%（2603 頁中 1577 頁、14860 文字中 8821 文字）が埋まっているが今後その拡充に努めたい。

[*15] TIFY はレスポンシブデザインを実現しており、さまざまな端末で快適に利用できる。

参考文献

[1] Tomohiko Morioka. Multiple-policy character annotation based on CHISE. *Journal of the Japanese Association for Digital Humanities*, Vol. 1, No. 1, pp. 86–106, 2015 年 11 月.

[2] 劉冠偉. 『大字典』和訓データベース構築の現状と課題. 情処研報, Vol. 2016-CH-110, No. 9, pp. 1–4, 2016 年 5 月.

[3] 高田智和. 『大字典』データベースをつくる. じんもんこん 2001 論文集, 情報処理学会シンポジウムシリーズ, 第 2001 巻, pp. 221–228. 情報処理学会, 2001 年 12 月.

[4] 石塚晴通, 池田証寿, 岡墻裕剛. 漢字字体規範データベースとその応用. 東洋学へのコンピューター利用 第 17 回研究セミナー, 全国文献・情報センター人文社会科学学術セミナーシリーズ、京都大学学術情報メディアセンター 第 78 回研究セミナー, pp. 53–63, 2006 年 3 月.

[5] 守岡知彦. Wiki 的手法に基づく構造化データの編集について. じんもんこん 2010 論文集, 情報処理学会シンポジウムシリーズ, 第 2010 巻, pp. 33–40. 情報処理学会, 情報処理学会, 2010 年 12 月.

[6] 守岡知彦, 劉冠偉, 高田智和. 漢字字体規範史データセット用従来型 UI 再生の試み. 情

処研報, Vol. 2019-CH-120, No. 2, pp. 1–6, 2019 年 5 月.
[7] 高田智和. 大字典データベースをつかう. 情処研報, Vol. 2004, No. 58 (2004-CH-62), pp. 45–52, 2004 年 5 月.

