4ZL-03

函館市史：統計史料編のデジタルデータ化に向けた校正作業の支援

菊村苑香† 渡部丈‡ 中小路久美代‡
公立はこだて未来大学† システム情報科学部‡

1. はじめに

紙媒体で保存されてきた統計史料をデジタルデータ化し，デジタルアーカイブとして利用したい．しかしながら，現時点で自動認識させたデータには，数値や文字が誤認識されている箇所が散見され，データとして使用することは難しい[1]．

本研究では，そういった誤認識箇所を訂正するための校正を行うユーザインタフェースの構築を目指す．函館市史:統計史料編[2]は，明治初期から昭和後半にかけての函館市史の統計史料表を編んだものである．これまでに PDF 及び JPEG としてデータ化されている．本研究では，JPEG 化された画像データから既存の文字認識技術を用いて，CSV 化させたデータを用いて行う校正作業支援に取り組む．

藤原[3]の研究では，函館市史:統計史料編のデジタル化に向けて，まず，全ページを個別画像としてスキャンしたデータを作成した．スキャンしたページの画像データに前処理を施して文字認識ソフトウェアに入力することで，認識精度を向上させることに成功している．

函館市史：統計史料編の誤り箇所を訂正する校正作業では，424 個のスキャン画像をそれぞれに対応するエクセルファイルのシートを結合した表データと比較して行う必要がある．試行実験として，シートを結合した 1 つの表データと，それが対応するスキャン画像を見比べて認識誤りを訂正してみたところ，人力でテキストファイルに記録するのに約 1 時間半を要した．この認識誤りの訂正を 424 回記録することはあまりにも膨大な作業となる．

本研究では，藤原らの研究成果から出力されたエクセルファイルのデータに対して，このような校正作業をより簡便に行えるように，元の画像と見比べながら，その誤り箇所を訂正するための校正作業を行うユーザインタフェース(UI)を構築することとした．

Support for Proofing Work for Digitization of Statistical Historical Materials of the City of Hakodate

†Sonoka Kikumura, Future University Hakodate
‡Jou Watabe, Future University Hakodate
‡Kumiyo Nakakoji, Future University Hakodate

2. 統計表認識データの校正作業

2.1 校正作業の試行

ランダムに選択したエクセルファイルのシートを対象として，エクセルファイルのシートを結合した表データと，元のスキャン画像を比較し，発見された誤認識の情報をテキストファイルに記録した．テキストとして記録するにあたっては，出力されたエクセルファイルの数値や文字が誤っているセルの位置（〇列△行）と，そのセルがどのように誤って認識されているか（文字や数値が異なる，全角を半角にするなど）を記録した．

2.2. 誤認識情報の構造化

誤認識箇所の記録を通して，誤認識情報として，エクセルデータ上の誤認識の箇所，および，誤っている文字や正しい表示，どのように誤っているかの説明，といった情報が必要であることが明らかとなった．

また，誤りの種類にはいくつかのパターン（型）があることが認められた．そこで，校正作業の試行を通して，誤認識情報の構造化を行った．構造化した各誤認識情報は，誤認識位置（列と行），誤りのパターン，備考欄（誤りの具体的な説明），現状のセルのテキスト内容，および，修正後のテキスト，フォントサイズ，割付，文字体形の 9 項目から成る．

3. 校正作業のためのユーザインタフェースのデザイン

統計資料編の誤り箇所の校正を簡便に行うことを目指し，Web ブラウザベースで校正作業のためのユーザインタフェースを構築することにした．校正作業のためのユーザインタフェースのデザインを図 1 に示す．

インタフェース右上にあるのは統計資料編の目次部である．目次部では，函館市史：統計史料編の目次に記載されている項目の一覧を表示する．目次部直下には，統計資料編の目次部で選択した項目に含まれる表の選択部がある．選択した項目に存在する表の一覧をサムネイル画像として表示する．

インタフェース左上は，スキャン画像とそれに対応する CSV ファイルを HTML における table として出力したものを表示する領域である．「Image」と書かれている下にあるのはスキャン画像表示部であり，「CSV」と書かれている下にあるのは CSV ファイル表示部である．CSV ファイル表示部において誤認識をもつセルを発見したらそのセルをダブルクリックする．クリックしたセルの背景色はオレンジ色に変化し，CSV ファイルの表の誤認識箇所が一目でわかる．

スキャン画像表示部にはオレンジ色の十字の線を，CSV ファイル表示部には緑色の十字の線を表示している．これらは誤認識の発見の補助をする役割を担う線である．元のスキャン画像と CSV ファイルを見比べて誤認識セルを探している内に，どのセルを見ていたかわからくなってしまう．こういった場面で補助線が有用であると考えられる．

スキャン画像表示部の右上にあるのは画像の拡大，縮小ボタンである．スキャン画像表示部に表示されている元のスキャン画像の拡大，縮小が可能であり，元のスキャン画像のサイズによって誤認識を発見しづらくなってしまうのを防ぐ．

スキャン画像表示部の左上にあるのは，統計資料編の本編の閲覧ができるボタンである．クリックすると函館市史：統計史料編の凡例の閲覧が可能である．

インターフェース左下にあるのは校正作業部である．校正作業部の左側では，CSV ファイル表示部においてダブルクリックしたセルの「位置」，「現状値」が同時に表示される．「訂正値」とかかれたテキストボックスには正しい値をユーザが入力する．訂正値の下の「備考欄」には，セルの中のどの文字がどのような誤りをしているかを入力することができる．

校正作業部の右側では，誤認識の型をチェックボックス内で選択する．校正作業の試行を通して同定した A 型から D 型のパターンに当てはまらない誤認識の型があった場合，「Add Pattern」とかかれたテキストボックスに型名を入力し「Add」ボタンをクリックすることで型の追加が可能である．「del」ボタンをクリックすると追加した型の削除が可能である．

校正作業部の左にある登録ボタンをクリックすることで，校正作業部で入力した校正情報がローカルストレージに登録され，CSV ファイル表示部で選択していたセルの背景色がオレンジ色から黄緑色に変化する．誤認識を修正したセルが一目でわかるようになっている．

CSV ファイル表示部の右上の Download ボタンを押下して，CSV ファイルがもつ誤認識情報を全誤認識が訂正された CSV ファイルをダウンロードできる．

4. システムを用いた校正作業と今後の展開

図 1 に示すシステムを用いた校正作業を登録する手順を説明する．ユーザは，統計資料編の目次部に書かれている項目から自分が校正作業をする対象の項目をクリックし，選択する．表の選択部に表示されたサムネイル画像の中から，校正作業を行う画像をクリックする．スキャン画像表示部にスキャン画像，CSV ファイル表示部に CSV ファイルが表示されたら，スキャン画像の拡大・縮小や補助線を活用して誤認識箇所を探す．誤認識箇所を発見したら，誤認識があるセルをクリックし校正作業部で訂正値，備考を入力する．誤認識の型を選択し，登録ボタンをクリックする．この作業を同様に繰り返し，表にある誤認識をすべて登録した後，Download ボタンをクリックすることになる．

今後は，本システムを用いた校正作業を全統計表に対して行う．誤認識データの修正版を作成すると共に，誤認識の傾向や特徴がわかれば，認識率向上にもつなげられると考えている．

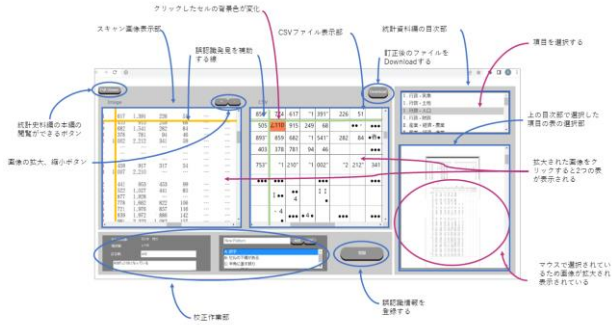

図 1:校正作業のためのユーザインタフェース

参考文献

[1] 中小路久美代，山本恭裕，松原伸人，川嶋稔夫，木村健一，函館市史：統計史料編のデジタルデータ化における多角的検討,情報知識学会誌，30(2)，176-181，2020.

[2] 函館市史編さん室(編)：函館市史：統計史料編，函館市，1987.

[3] 藤原慎太郎，印刷された統計史料表のテキスト化技術の開発，卒業論文，公立はこだて未来大学システム情報科学部，2022.