Vol.2010-MPS-81 No.35
Vol.2010-BIO-23 No.35
2010/12/17

近代デジタルライブラリーテキスト化支援のためのポータルサイトの設計

黒田佳世[†1]　榎本友里枝[†1]
高田雅美[†1]　城和貴[†1]

本稿では，本の見開き画像をテキスト化する際に生じる誤認識を修正する機能をもつポータルサイトを設計する．対象とする画像は，国会図書館が所有する近代書籍のデジタル画像とする．このデジタル画像に対して，近代書籍専用の活字文字認識を適用することによりテキスト化は可能であるが，誤認識が生じる．これを修正するために，デジタル画像とテキスト文書を用いる．提案するポータルサイトでは，テキスト文書と画像データを対応させるメタデータを作成することによって，テキストに対応する画像を表示させることを可能にする．

Design of a Portal Site for Textizing Early-Modern Printed Books

Kayo Kuroda,[†1] Yurie Enomoto,[†1] Masami Takata[†1]
and Kazuki Joe[†1]

In this paper,we present a design of a portal site which has functions for correcting erratums.These erratums are caused when the facing images,which are owned by Digital Library from Meiji Era in National Diet Library,are transformed into text documents.Although those images can be transformed by using an OCR,which is specialized in early-modern printed books,the OCR infrequently cause erratums.So,to correct them, we take by means of those images and text documents. In this case,we make the metadata by which text can be corresponded to image .

†1 奈良女子大学大学院人間文科研究科
Graduate School of Humanities and Sciences, Nara Womens's University

1. はじめに

現在，国立国会国立図書館では，希少価値のある書籍を経年劣化や人の手による破損・損失の危険を避け，一般向けに公開するために，書籍をページごとにマイクロフィッシュ化して，それをフィルムにスキャンした画像を Web で公開する近代デジタルライブラリー[1)] を開発している．この近代デジタルライブラリーでは，明治・大正期刊行図書約 17 万冊の資料本文をデジタル画像で閲覧できる．ただし，本文に関しては，データ形式が画像のみであり，テキストデータとして存在していないため，全文検索などの通常のテキストデータを扱う際に利用できる機能には対応していない．そのため，近代書籍データのより簡便な利用のために早急なテキスト化が望まれている．

一般的な文書画像であれば，OCR によってテキストデータへの変換を自動で行うことができる．これによって作成されたインターネット上の書籍閲覧・検索システムとして，Google ブック[2)] がある．しかし，近代書籍では，現在使われていない古い日本文字や表記の利用，使用されている活字も出版社や出版時期によって異なる等の問題があるため，市販の OCR は正常に認識することができない．そこで，近代書籍に特化した OCR システムを作成し，画像データをテキスト文書化するという試みがなされている[3)4)]．しかし，一般的に市販の OCR でも誤認識が生じるのと同様，開発中の OCR を用いたテキスト文書には誤りが生じる．

例として，近代書籍に特化した OCR を用い 10 種類の文字のデータを対象とした場合の精度は 97．8 %という結果を得ている[4)]．また，日本語の文字コード体系は JIS によって，第 1 水準漢字，第 2 水準漢字，非漢字などが規格化されている．この中で，第 1 水準漢字，第 2 水準漢字で制定されているのは 6355 文字の漢字である．この 6355 文字の文字データを対象として OCR を用いた場合の精度は，文字の種類が増えるにつれて下がることが予想される．

このような OCR によって生じるテキスト文章の誤りをユーザが修正できる機能は，Google ブックにも存在しない．そこで本稿では，テキスト文書閲覧者であるユーザが書籍文章を閲覧中に誤りを見つけた際に，そのテキスト文書の位置に対応する元の画像データを表示し，テキスト修正画面を表示させ，誤りを訂正する機能を持つポータルサイトの設計を行う．対応する画像の表示にはメタデータを用いる．さらに，テキスト文書とそのデジタル画像を照らし合わせたあとで，正しい文書を理解し，誤りを修正する機能もメタデータを用いることにより行う．なお 現時点ではテキスト化が未完了であるため，テキスト文書は青空文庫[5)]

© 2010 Information Processing Society of Japan

Vol.2010-MPS-81 No.35
Vol.2010-BIO-23 No.35
2010/12/17

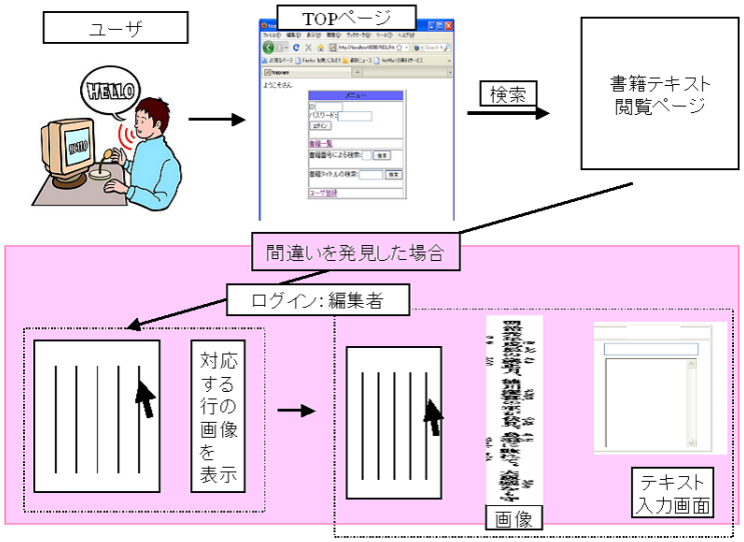

図 1 ポータルサイトの概略図

で提供されているテキストを使用する．

2 章ではポータルサイトの機能について述べ，3 章でデータベース，4 章で画像表示とテキスト修正法ついて説明し，最後にまとめを述べる．

2. ポータルサイトの機能

近代デジタルライブラリーで提供されている画像は，明治・大正刊行期の書籍をマイクロフィッシュ化してそれをフィルムスキャンしたものであるため，ノイズを完全に除去することは困難である．また，OCR を用いた文字認識において，誤認識が存在する．このノイズや OCR で生じた誤認識を自動的に修正することは非常に困難である．そこで，ユーザにより誤りを修正できるポータルサイトが必要であると考えられる．

以降，id を持ちテキストを編集できるユーザを編集ユーザ，テキストを閲覧するのみで編集はしないユーザを利用者，管理権限を持つユーザを管理者，編集ユーザと管理者を編集者，編集ユーザと利用者と管理者を総称してユーザと定義する．

図 1 は提案するポータルサイトの概略図である．ユーザは近代書籍に特化した OCR を

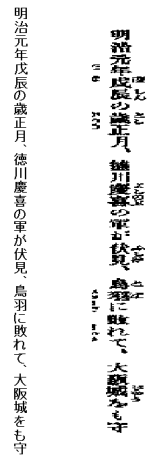

図 2 1 行のテキストと画像

用いてテキスト化されたテキスト文書を Web で閲覧することができる．その際，テキスト文書に誤字脱字を見つけた場合に，編集者はユーザ id とパスワードを用いてログインする．その後，図 1 右下の画像データのようなテキスト文章に対応した画像を閲覧できるようにする．このとき，表示される画像データはユーザが指定したテキストの 1 行のみの画像データとする．**図 2** は，実際の対応画像 (1 行文) の画像表示とそれに対応する文書である．この画像データと, テキスト文書とを照らし合わせることで正しい文書を理解する．そして，テキスト文書の修正という機能を用いて誤りを修正する．これは，テキスト文書を修正できる編集者をある程度規制し，悪意的なテキストの変更などを避けるためである．なお，書籍検索の段階では，ログインの必要性はない．画像データで表示される画像は，JPEG2000 や JPEG 形式の画像データを pbm 形式に変換したものである．pbm 形式に変換して利用する理由は，画像の容量の減少と画像処理での扱いやすさである．

GUI の開発環境として，サーバに Apache Tomcat6．0 をインストールし，ブラウザは Firefix3．6 を使用する．Firefox をブラウザに選択理由は Top 画面に使用しているフレームがエクスプローラは非対応だったためである．また，インターフェースと検索プログラムを開発するにあたり Java を適用する．Java は XML ファイルを検索するためのパーサも標準で備えているので，別途 XML パーサを作成する手間を省くことが可能である．さらに，

© 2010 Information Processing Society of Japan

Vol.2010-MPS-81 No.35
Vol.2010-BIO-23 No.35
2010/12/17

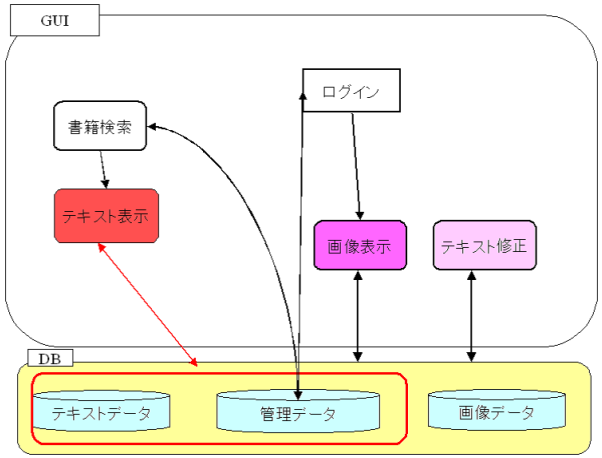

図 3 システム全体図

JSP やサーブレットを利用することで動的に HTML などを出力させることができるため検索結果などを容易に表示することが可能である．

ポータルサイトは画像の閲覧・テキストの修正のほかに書籍の検索機能を持つ．検索方法は 2 通りである．1 つ目は，書籍のタイトルによる検索，2 つ目は書籍番号を用いた検索である．一般的には，書籍のタイトル，著者による検索が行なわれるが，近代デジタルライブラリーが提供している csv データには書籍検索に用いるデータに著者が含まれていないためこの 2 つを用いる．なお，データの詳細については次章以降で記述する．**図 3** は，システムの全体図である．図 3 からわかるようにユーザが行える動作は，書籍検索，テキスト表示の 2 つのみである．編集者が行なえる動作は，ログイン，書籍検索，テキスト表示，画像表示，画像修正の 5 つのみである．これらから得られる情報を基に，開発するポータルサイトが自動的にデータベースにアクセスし結果を表示する．

テキスト修正では，編集者が 1 行分のテキストを入力し，更新ボタンをクリックすることによりテキストが更新される．打ち込みを行っただけでは修正は完了せずこの更新ボタンを押して初めて更新が反映される．近代書籍において，1 行の文字数が 100 文字を超えることはない．そのため，悪質な書き換えをさけるため入力文字は 100 文字以下という規制を 設ける．この規制は，100 文字以上入力された場合にエラー文を表示させることにより行う．

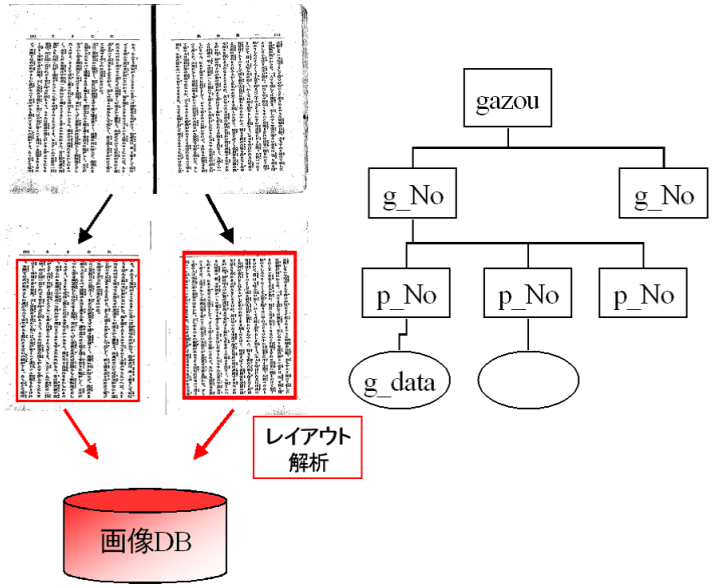

図 4 画像 DB

また，複数のユーザが同時に修正することを防ぐためにトランザクションの処理を行っている．これは，修正機能を使用する場合にその他のユーザは修正できないようロックをかけるという仕組みである．この処理を有効に使用するためには，ロックする時間の制限も必要になってくるためタイムアウト機能を利用する．

3. データベース

ポータルサイトで使用するデータは主に 3 つのデータにわけることができる．画像データ，テキストデータ，管理データである．以下で 3 つのデータについて述べる．

画像データとは，近代デジタルライブラリーで提供されている JPEG2000，もしくは JPEG 形式の画像をノイズ除去などの前処理を行い，さらにレイアウト解析を行なった後の画像である．この画像データは，ポータルサイトの所有するデータ容量の大半をしめている．そこで，容量を減らすために，JPEG2000 や JPEG などの形式から pbm 形式に変換

© 2010 Information Processing Society of Japan

Vol.2010-MPS-81 No.35
Vol.2010-BIO-23 No.35
2010/12/17

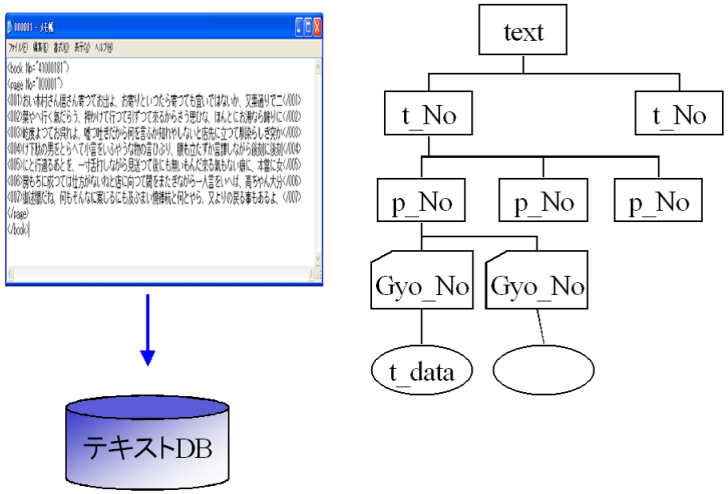

図 5 テキスト DB

する．画像データは見開きの 2 ページのまま使用せず，分割して半ページにし，ノイズ除去を行ない，レイアウト解析して取得したデータを扱う．その理由は，画像データの画像の容量削減と座標解析の効率化のためである．見開きのページは，ページ間の影が入るためこれを文と認識しないために取り除く必要がある．さらに，半ページのデータのままでは容量が大きいため，ノイズ除去を行い，レイアウト解析で絵や図などを除き文章の範囲のみを抽出する．また，画像はさらに 1 行ごとや文字ごとに分割しない．これは，1 行ごとや文字ごとに画像を分割してもこれ以上のノイズ除去などによるデータ容量の削減はできないためである．さらに，画像へのファイルパスなどの管理情報は増えることにより，ポータルサイトのデータを増やすことになるため効果的ではないと考えられる．よって，画像データのファイル構成は**図 4** の通りである．

次にテキストデータについて述べる．本来のテキストデータは OCR を用いてテキスト化されたデータを使用する．画像データ半ページごとにテキストファイルを作成する．**図 5** は左がテキスト DB に格納されるテキストファイルの一例である．右は，テキストファイルに付与するメタデータタグの構造である．テキスト文書には，ます書籍番号（t_No）とページタグ（p_No）を付与する．そして，それぞれの行に行番号の XML タグ（Gyo_No）

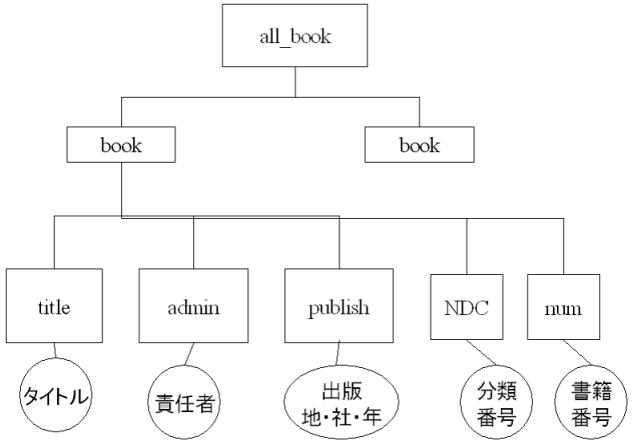

図 6 csv ファイルのデータの構成

を付与する．この時，番号は 001 のように 3 桁で記述する．3 桁の番号を使用する理由は，近代書籍の大部分は，1 ページに含まれる行数が 2 桁であるため，より万全を期して 3 桁を用いている．このテキストデータを用いることによって，検索の際に行う解析やソートの簡単化を可能とする．

次に管理データについて説明する．管理データは大きく 3 つの形式に分けられる．csv ファイルとテキストファイルと XML ファイルである．

まず csv ファイルについて述べる．**図 6** は，csv データの構成である．この csv ファイルには含まれない書籍が存在する．この書籍については本稿では自作して設計を行う．本稿で使用した csv ファイルは近代デジタルライブラリーで提供されているファイルである．このファイルには，近代デジタルライブラリーで提供されている書籍のタイトル，責任表示，出版地，出版社，刊行年，NDC，全国書籍番号が格納されている．このファイルを使用する利点は，た新に書籍情報用のデータを作成する必要がないことと，csv ファイル形式で提供されているため DB での検索が容易であるということである．

次に，テキストファイルについて述べる．このファイルは，編集者がログインする際のユーザ認証を行うときに使用されるファイルである．また，このファイルはユーザは閲覧や書き込みなどはできない．ユーザ認証する際に必要となるパスワードの暗号化の仕組みは国会図書館の仕様を用いる．

© 2010 Information Processing Society of Japan

Vol.2010-MPS-81 No.35
Vol.2010-BIO-23 No.35
2010/12/17

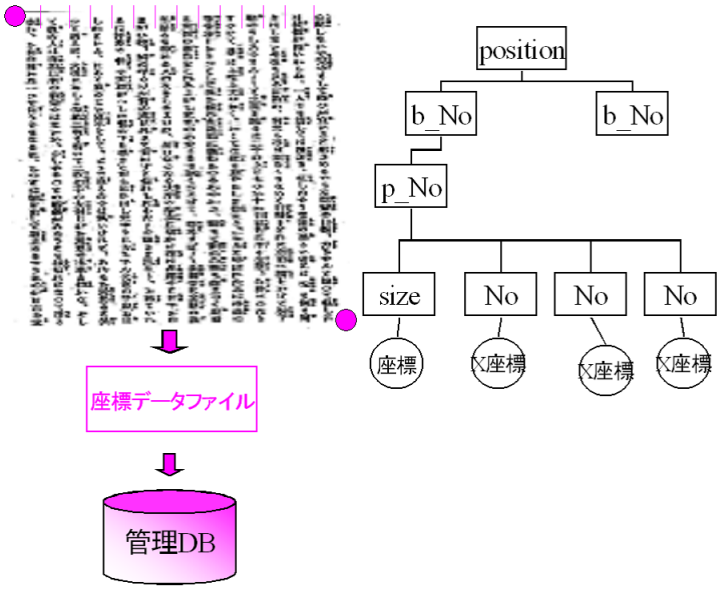

図 7 座標解析

3 つめの XML ファイルについて述べる．XML ファイルには，画像の座標メタデータと詳細 XML がある．この 2 つの XML ファイルは統合することもできるが，解析の効率化や時間短縮の理由により 2 つに分割する．[?)?)] まず，画像の座標メタデータの XML ファイルについて説明する．最初に，**図 7** のようなレイアウト後の画像データに対し，画像の左上の座標値と右下の座標値を求め，ファイルに記述する．次に，図 7 の行間の線のように，各行ごとに行間の閾値をもとに行間の x 座標を求める．この求めた座標を XML ファイルに記述し，各行番号のタグを付与させる．このときの番号は 1 行目であれば，001 と記述し解析を簡単化できるようにする．次に，詳細 XML ファイルについて説明する．この詳細 XML は，**図 8** の構成のようにテキストデータ，画像データ，画像解析データ，csv データすべてへのファイルパスを保持する INDEX の役割をはたす．これは，実際にユーザがテキストの閲覧以外に画像の閲覧やテキストの修正をする際に図 7 の解析 D のように，XML にアクセスし，テキスト，画像，座標を一度にまとめて扱えるため，画像解析情報と画像情報を同時に扱える利点がある．

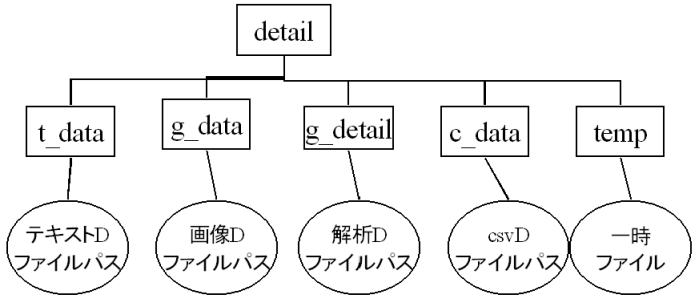

図 8 詳細 XML

4. 画像表示とテキスト修正法

画像表示とテキスト修正のプログラムの流れについて述べる．

まず，画像表示について説明する．**図 9** は画像表示のプログラムの流れを表している．画像表示の選択がされた場合，最初に，管理データベース内の詳細 XML にアクセスし，画像データと座標解析データのパスを取得する．次に，テキストファイルの書籍番号とページ番号のタグ情報から，画像データと座標解析データそれぞれの該当するファイルを呼出す．次に，座標解析ファイルより画像の左上と右下の座標（図 7 の左の丸点）を取得し，画像全体のサイズを取得する．そして，表示する行を挟む 2 つの X 座標（図 7 の左の X 座標）をテキストファイルの行番号を用いて取得し，この X 座標と全体の画像データを用いて指定した 1 行分のみの画像表示を行う．

次に，テキスト修正プログラムの流れについて説明する．**図 10** はテキスト修正のプログラムの流れを表している．テキスト修正の場合，必要となるのはテキストデータと一時ファイルである．そのため，最初に管理データベース内の詳細 XML にアクセスする．テキストファイルのパスを取得したら書籍番号・ページ番号のタグを用いてテキストファイルを検索し，呼び出す．次に，行番号タグより指定された行の位置を取得し，その部分のみ上書きし，保存する．また，一時ファイルには修正を行なったユーザ名と修正を行った時間を書き込んで保存しておく．このように，テキストを修正したユーザ名と更新した時間をデータとして保持しておくことで，管理者がそれらを用いて訂正したユーザ情報を取得でき，また，

© 2010 Information Processing Society of Japan

Vol.2010-MPS-81 No.35
Vol.2010-BIO-23 No.35
2010/12/17

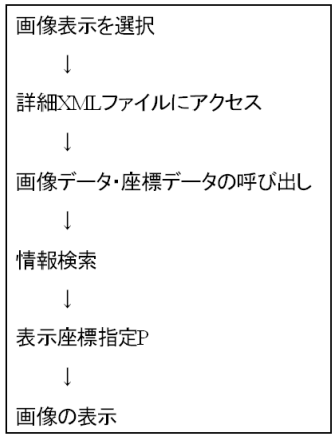

図 9 画像表示の流れ

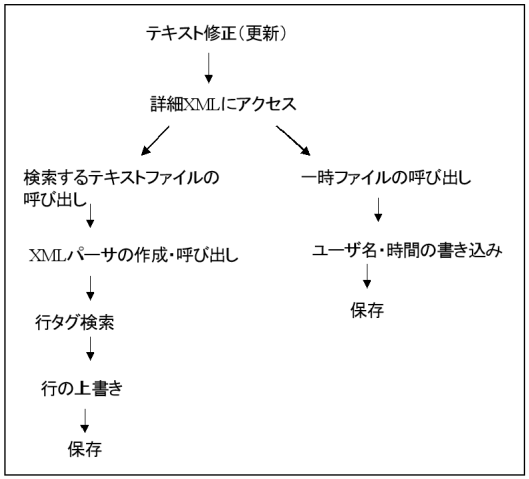

図 10 テキスト修正の流れ

誤りの頻度の解析なども可能にする．さらに，このファイルにアクセスしたり，データを削除したりできるのは管理者のみである．これらの処理は，悪意的な書き込みなどを防ぐことも可能にする．

5. おわりに

本稿では，使用されている活字が出版社や出版時期によって異なる多フォントの OCR の誤認識をユーザが元の画像を使用することにより修正するためのポータルサイトの設計を行った．このポータルサイトを使用することにより，ユーザは誤認識を見つけたときに，対応する画像を表示させることができる．また，表示された画像とテキストと照らし合わせることにより間違いを理解し，行ごとの修正もすることができる．

今後の課題として，表示される画像が行ごとであり，かつ，pbm 形式の画像であったために文字の読みにくさが感じられたため，JPEG2000 や JPEG から pbm 画像への変換の際の画像情報の劣化を防ぐためのプログラムの向上や GUI での文字拡大機能を新たに追加するなどの改良をするべきである．

参 考 文 献

1) 近代デジタルライブラリー :http://kindai.Ndl.go.jp/ .
2) Google ブック検索：http://books.google.co.jp/books/ .
3) 芦田尚美 高田雅美 木目沢司 城和貴：近代書籍に特化した多フォント活字認識法, 情報処理学会研究報告, MPS-73，Vol.2009，No.19，pp.205-208 (2009.2)
4) Ishikawa, C. , Ashida, N. , Enomoto, Y. , Takata, M. , Kimesawa, T. , and Joe, K.:Recognition of Multi-Fonts Character in Early-Modern Printed Books, Proceedings of International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA' 09)
5) 青空文庫：http://www.aozora.gr.jp .
6) 豊島良美, 石川千里, 高田雅美, 城和貴：短期地震研究のための統合的なマルチデータベースの設計，情報処理学会研究報告，vol2006, No135, pp. 65 ー 69 (2006.12).
7) 豊島良美, 石川千里, 高田雅美, 長尾年恭, 城和貴：地震短期予測研究のための地電流解析ポータルの開発，情報処理学会研究報告，vol2009, No19, pp. 209–212 (2009.2).

© 2010 Information Processing Society of Japan