ブックバーコーディング法：版本の差読に基づく「武鑑全集」の網羅的な解析に向けて

北本 朝展（ROIS-DS 人文学オープンデータ共同利用センター，国立情報学研究所）
藤實 久美子（国文学研究資料館）
本間 淳（フェリックス・スタイル）

概要：本論文は版間差分を読む「差読」という方法論を実現するアルゴリズムである「ブックバーコーディング法」を提案する．そして，画像位置合わせと安定結婚問題のアルゴリズムを組み合わせることで，版本の網羅的な照合が可能なことを示す．また，この成果を「武鑑全集差読プラットフォーム」として公開し，これを活用した差分翻刻の試行を通して，提案手法の有効性を確かめた．提案手法は任意の版本に適用できるため，将来的には汎用の版本照合研究基盤に発展する可能性がある．
キーワード：版本照合，差読，武鑑，画像位置合わせ，安定結婚問題、ブックバーコーディング

Book Barcoding Method: Comprehensive Analysis of the Bukan Complete Collection using Differential Reading of Woodblock-Printed Books

Asanobu Kitamoto (ROIS-DS Center for Open Data in the Humanities / National Institute of Informatics)
Kumiko Fujizane (National Institute of Japanese Literature)
Jun Homma (FLX Style)

Abstract: This paper proposes the “Book Barcoding” algorithm to realize the “differential reading” methodology for reading the difference between versions of woodblock-printed books. First, we show the possibility of a comprehensive collation of woodblock-printed books by combining the algorithms of image registration and stable marriage problem. We then confirm the effectiveness of the proposed method through the experiment of differential transcription using the “differential reading platform” for the Bukan Complete Collection. Finally, the proposed method can be applied to any woodblock-printed books, having the prospect of evolving into the general-purpose collation platform for woodblock-printed books.
Keywords: Collation of Wooldblock-Printed Book, Differential Reading, Bukan, Image Registration, Stable Marriage Problem, Book Barcoding

１．まえがき

本論文は木版印刷に基づく版本を照合する方法論である差読（differential reading）のための技術基盤の確立について述べる．デジタルヒューマニティーズ研究では，機械の支援を受けることで，人間による精読（close reading）とは異なる読み方を開拓する研究が盛んである．例えば遠読（distant reading）は，テキストマイニングなどの支援によって，テキスト集合の統計的な性質を読んでコーパスの全体像を解釈する．同様に本論文が扱う差読は，コンピュータビジョンなどの支援により，バージョン間の差分（版間差分）という元の版本にはない情報を解釈することを目指す．

木版や活版による印刷本を照合し，異なる版で生じた異同を特定する作業は，誤りの修正や情報の更新，テキストの時間順序の推測など，多くの研究における基本的な作業として汎用性が高い．こうした差分をテキストの比較から検出する研究はこれまでも数多く行われてきたが，本論文は画像の比較から版間差分を検出するという新しい方法論を提案する．そしてこの方法論を最初に適用する対象として「武鑑全集」を選んだ．武鑑は 200 年間にわたって内容を更新しながら出版が続いたことが知られており，版間差分の研究に最も適した資料の一つである．

筆者らはこれまで，武鑑という「歴史ビッグデータ」を網羅的に分析するための技術基盤の構築に関して，数編の論文を発表してきた．武鑑全集プロジェクトの問題設定を整理し[1,2,3]，ページ照合のアルゴリズムを開発し[4,5]，さらにブック照合のアルゴリズム[6]を開発した．本論文はこうした過去の成果を踏まえ，アルゴリズム相互の関係を整理し，それらを総合した概念としての「ブックバーコーディング法」を提案するものと位置付けることができる．

本論文は，第 2 章で武鑑全集を紹介し，第 3 章で差読の方法論を整理する．次に第 4 章と第 5 章ではページ照合およびブック照合のアルゴリズムを概観する．さらに第 6 章は武鑑全集に対する差分翻刻の試行をまとめ，第 7 章は汎用的な版本照合研究基盤への発展の可能性を論じる．

2.「武鑑全集」と時系列データ構造化

2.1 武鑑全集とは

武鑑とは，江戸時代に出版された大名家および幕府役人の名鑑である．武鑑とは 17 世紀中ごろに出版されはじめ，慶応 3 年（1867）10 月 14 日の大政奉還まで，200 年以上の間，出版され続けた．武鑑は実用書であり，ロングセラーブックであった．武鑑は，社会の需要に応えて，年を追うごとに厚くなり，その改訂の頻度は年に数度から月に数度までに増えた[7]. このように江戸時代を通して発行され続けた武鑑であるが，これが歴史研究の面でも興味深い理由を以下にまとめる．

第一に武鑑は江戸時代のデータブックであり，多種類の情報や多数の人名などがまとまっているという点である．このことは幕藩体制に関する細粒度データとして価値が高い．第二に武鑑の更新頻度が高いため，時系列データとして価値が高い．例えば武鑑上で人物情報の変化を追跡すれば，人物や役職の変動などを定量的に把握できる可能性がある．第三に項目ごとの整理がされたデータであるため，構造化データとして価値が高い．大名の参勤交代時期や献上品などはすでに項目ごとに整理されているため，現代の我々が改めて整理し直す必要はない．第四に分量が膨大であることから，ビッグデータとして価値が高い．すべての武鑑をテキストとして翻刻することは不可能であることから，歴史ビッグデータ研究のための構造化データを構築するには新しいアプローチが必要である．

このように武鑑を歴史ビッグデータとして活用するをための研究プロジェクトが「武鑑全集」[8]である．ここでは，国文学研究資料館（NIJL）がデジタル化し，ROIS-DS 人文学オープンデータ共同利用センター（CODH）が公開する，381 点の武鑑のデジタル画像を公開している．この大規模なデータの分析手法として，我々は共時的な分析と経時的な分析という 2 方向の研究を同時に進めている．

2.2 共時的な分析

武鑑における共時的な分析とは，時間（＝版）を固定し，大名家などの実体を変化させることで，ある時間断面のデータを網羅的に構造化し分析する手法を指す. 武鑑全集の場合は，『寛政武鑑』（寛政 1 年, 1789）[9]を基準として選び，この版本のデータ構造化を進めている．現在までに作成した構造化データは以下の通りである．

1. 大名家に関連する基礎情報（名称，石高等）
2. 大名家に関連する地理情報（居城地，上屋敷，菩提寺）
3. 参勤交代に関する情報
4. 紋／道具に関する情報
5. 献上品／拝領品に関する情報

これらの項目に関する構造化データを大名家 264 家ごとに作成し，外部リソースである「江戸マップ β 版」[10]などともリンクすることで，資料横断的な大名家の分析を可能とした．現在のところデータの構造化は以下の手順で進めている．

1. IIIF Curation Platform [11]を用いて，武鑑の各データ項目を含む領域を指定し，キュレーションに追加する．
2. 各領域に含まれる文字を翻刻し，データ項目ごとにエクセルに書き込む．
3. エクセル表データをパースし，キュレーションと対応付けるなどの処理を行って，「武鑑全集」ウェブサイトに展開する．

このように IIIF Curation Platform とエクセルを組み合わせることで，データと出典がリンクする形式で構造化データを作成した．その他，武鑑全集ウェブサイトの工夫にも触れておきたい．

1. 地理情報が付与できるデータは，Google Maps 上にも表示する．
2. 参勤交代データは，江戸との往復を行う時期が明示されているため，d3.js を用いたアニメーション地図として可視化する．
3. 紋／道具から色情報を取り出し，色名に対応する RGB 情報と共に表示する．

2.3 通時的な分析

武鑑における通時的な分析とは，大名家などの実体を固定し，時間（＝版）を変化させることで，ある実体断面のデータを網羅的に構造化し分析する手法を指す．これは大名家に関する「時系列データ」を構築し，江戸時代 200 年間にわたる変化を網羅的に分析し，新たな歴史事実を見出すことを目指した研究のための作業である．例えば，江戸時代の出版業は情報更新にどれだけ素早く反応したのか，大名家や幕府役人の役職交代やキャリアパスはどうだったのかなど，このデータは新たなリサーチクエスチョンを調べる土台となる．さらに武鑑に登場する人物と柳営補任など他の資料に登場する人物とを相互に検証して江戸時代の人物データベースを構築できれば，江戸幕藩体制に関する歴史ビッグデータ研究に新たな可能性が開けるだろう．

こうした時系列データを効率的に構築するには，ある版と別の版とを照合し，どの部分の情報が更新されたかを容易に確認できる仕組みが欲しい．この課題に対して本研究が提案するのが「差読」，つまり版間差分を強調または抽出することで，多数の版を翻刻する人間の労力の軽減を目指す方法論である．

3. 差読の方法論

3.1 テキスト照合と画像照合

差読という方法論については，すでに過去の論文[1,2,6]でも述べてきたが，本論文でも重要な部分に絞って改めて議論しておきたい．差読とは，

機械の支援を受けて版間差分を強調することで，版間に生じた変化を読みやすくする方法論である．本論文では，これをテキストの照合ではなく画像の照合によって行うが，それは武鑑という資料の特徴に理由がある．

第一に，武鑑には多数のバージョンが存在して分量が膨大なため，全体を翻刻することは現実的でないからである．第二に，武鑑はデータブックとして全体が表組のレイアウトであるため，テキスト化の方法を一意に定めることが難しいからである．第三に，紋などのグラフィック要素の更新や木版の欠損などの非文字情報の差読は画像の照合でないと実現できないからである．文字の点画の相違よりも匡郭や界線の欠損の有無の方が，板木の前後関係を推測する有力な手がかりとなる場合もあり，画像の照合に基づく差分検出は版本研究でも重要な位置づけを占めている．

なお板本書誌学[12,13]では，(1) 刊（板・版），(2) 印（刷・摺），(3) 修（補・訂）を区別する．「刊」とは新しい板木を彫って本を刊行すること，「印」とは既存の板木を使って本を刷ること，そして「修」は既存の板木に対して埋木などを使って部分的な修正を加えることを指す．ソフトウェアのバージョン管理になぞらえれば，「刊」はメジャーバージョン，「修」はマイナーバージョンとみなせるかもしれない．

ここで特に注目すべきなのが「修」である．版（板）権という言葉に端的に表現されるように，板木は財産と位置づけられる高価なものであるため，たとえ修正が必要でも作り直しはせず，埋木を行うなど最低限の変更で対応することが多かった．そのため，板木全体としては類似しているものの，細部が異なる多数のマイナーバージョンが生み出される．このような「修」の検出とともに，「修」を越えた修正である「刊」を版間差分として検出することが，画像照合の目標となる．

3.2 画像照合とコンピュータビジョン

2 枚の画像を照合するという問題に対して，これまでは複数の資料を横に並べ，人間による目視で比較するという原始的な方法が使われてきた．しかし目視比較は短期記憶を酷使するため，人間にとっては負担が大きいタスクである．まさに「間違い探し」というゲームになるほど，人間にとっては難しく精度も低いタスクなのである．そうであるなら，このタスクを機械に任せることはできないだろうか．

その可能性を検討するため，2 枚の画像を照合するタスクを，2 枚の画像を位置合わせするステップと，2 枚の画像の差分を強調または抽出するステップに分解してみよう．前者はコンピュータビジョンの分野で画像位置合わせ（image registration）と呼ばれる問題に相当する．これは応用範囲が広い基本問題であることからすでに膨大な研究成果があるため，既存のアルゴリズムで十分に実現できる．一方，後者については，可視化によって差分を強調する方法と，物体検出や画像分類によって差分を抽出する方法の 2 つの可能性がある．後者の方がより高度な自動化を含む手法であるが，これは実は機械にとっては難しいタスクとなる．版間差分の中から，意味のあるシグナルと意味のないノイズを分類するには，版本に関する背景知識が必要となるため，精度を高めることが難しくなる．例えば匡郭や界線の欠損をシグナルと認めるのは難しい．一方，人間にとってこれはそれほど難しいタスクではない．シグナルの認識は短期記憶も必要なく瞬時にできるため，むしろ人間が得意とするタスクと言える．

ゆえに本論文の差読は，画像の位置合わせと差分の強調を機械が行い，差分の抽出は人間が行うという，細粒度の人機分業としてワークフローを構築する．全体としてみれば，画像照合という困難なタスクを簡単なタスクに反転させることで，全体のワークフローを効率化できる．

3.3 差分翻刻

差読のアプリケーションとして期待されるのが差分翻刻，すなわち 2 つの版を比較して「修」または「刊」が検出された部分のみを翻刻することで，不要な翻刻を省略するという新しい翻刻手法である．江戸時代の版本において「修」や「刊」の頻度がどのぐらいかという問いは，それ自体が江戸期の出版産業を考える上で興味深い研究課題であるが，少なくともその頻度に相当する分量の翻刻を省略できるはずである．また差分翻刻は作業量の削減にとどまらず，翻刻品質の向上にもつながりうる．版間差分によって複数の版の比較が可能となるため，より多くの情報に基づき翻刻を進めることができるからである．差分翻刻の試行については第 6 章にまとめる．

3.4 照合アルゴリズム

差読はブック照合とページ照合という 2 つのアルゴリズムの組み合わせとなる．まずブック照合とは，2 点の版本（ブック）の照合に基づき，同一板木から印刷されたページの組を判定するアルゴリズムである．一方，ページ照合とは，2 枚の画像（ページ）の照合から同一板木の異同を強調するアルゴリズムである．

2 つのアルゴリズムは，ブック照合の内部からページ照合が繰り返し呼び出されるという関係となる．ページ照合が実行するのは，3.2 節で述べた画像位置合わせアルゴリズムであるが，これをブック照合から見た場合，どの画像を照合する

ようにページ照合に指示するかという選択が必要になる．というのも，ページ照合の目的は，任意の 2 枚の画像を照合することではなく，同一の板木に由来する 2 枚の画像を照合することだからである．ところが，どのページがどの板木に由来するかというメタデータは，元の版本に存在しないという問題がある．そこで，同一板木に由来する 2 ページのペアを判定するという「同一板木判定」がブック照合の課題となる．そこで，第 4 章ではページ照合のアルゴリズムを説明し，第 5 章ではブック照合のアルゴリズムを説明する．

4．ページ照合

4.1 画像の位置合わせ

ページ照合で用いる画像の位置合わせには多くの手法があるが，大別すると線形変換と非剛体変換の 2 種類に分けられる．線形変換では画像全体を射影変換するが，非剛体変換では局所的な変形をつなぎ合わせた変換を考える．これを差読に当てはめてみると，カメラの撮影方法による影響は（レンズの歪みを除けば）線形変換の範囲で扱えるが，紙面の湾曲を扱うには非剛体変換が必要となる．本論文は，単純かつより高速なアルゴリズムとして，線形変換を対象とする．

線形変換のパラメータは 2 段階で求める．まず個々の画像から特徴点を抽出する．そのために，コンピュータビジョン分野の代表的ライブラリである OpenCV で提供される特徴点検出アルゴリズムを比較したところ，AKAZE 特徴検出器[14]の性能が最も高かったため[5]，本論文ではこれを用いて特徴点と AKAZE 特徴量を抽出する．抽出したデータは，リレーショナルデータベース（RDB）である PostgresQL に保存しておく．

次に 2 枚の画像を照合する際には，それぞれの画像から抽出した特徴量を RDB から読み出し，特徴量の距離に基づき特徴点を対応付け，射影変換行列を推定するという流れになる．特徴量の距離は，AKAZE 特徴検出器で標準的な距離であるハミング距離を利用し，射影変換行列の推定には標準的なアルゴリズムである RANSAC [15]を用いる．そして画像照合の品質指標として正対応（inlier）特徴点を数え，正対応特徴点の数が多いほど照合品質が高いとみなす．最後にこれらの推定結果も RDB に登録する．図 1 はページ照合のための例を示す．2 枚の画像で特徴点が対応する．

4.2 画像比較ツール vdiff.js

上記のアルゴリズムでは，ページ照合の結果は射影変換行列として得られる．しかし差読で欲しいのは行列そのものではなく，行列を基に変換し位置合わせした 2 枚の画像の方である．そこで位置合わせした状態で 2 枚の画像を表示する機能をウェブブラウザ上で実現するために，ウェブページ埋め込み型のオープンソース画像比較ツール vdiff.js を開発した[6,16]．開発言語は JavaScript であり，内部では OpenCV.js を利用している．

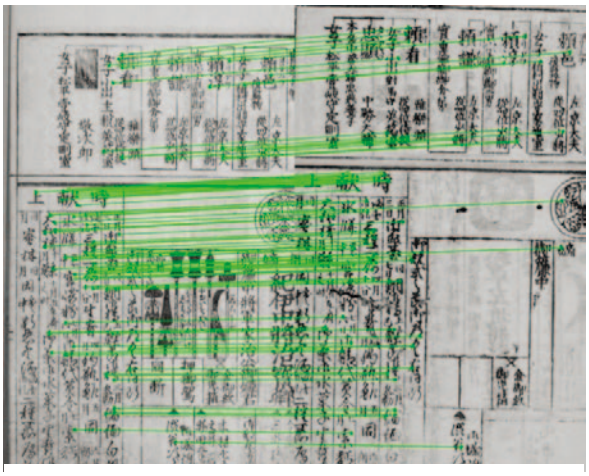

図 1：ページ照合のための特徴点マッチング．緑の線は正対応（inlier）特徴点の対応を示す．

Vdiff.js は射影変換機能を備え，画像内で比較したい対象の配置の違いを軽減して比較できる点に独自性がある．また vdiff.js editor は vdiff.js における重ね合わせ結果が満足できない場合，対応点を修正する機能を提供する．Vdiff.js は画像比較の用途によって，「局所強調比較モード」と「全体変化比較モード」という 2 つのモードを使い分けることができる．「局所強調比較モード」は同一出版物の異なる版の画像比較などでの利用を想定し，左右表示，強調表示，赤青表示という 3 種類の表示方法を提供する．一方「全体変化比較モード」は古写真と現代写真の今昔比較や災害写真のビフォー／アフター比較などでの利用を想定し，左右表示と透明表示という 2 種類の表示方法を提供する．武鑑全集では「局所強調比較モード」を利用する．

なお vdiff.js を呼び出す際に指定するパラメータは 4 つの対応点の組である．一方，画像照合から得られるパラメータは射影変換行列であり，そのままでは API の仕様に適合しない．そこで vdiff.js の API に与える 4 点を選ぶアルゴリズムとして，正対応特徴点の凸包上で最も離れた 4 点を選ぶというアルゴリズムを新たに開発した．

4.3 差読のための画像照合サービス

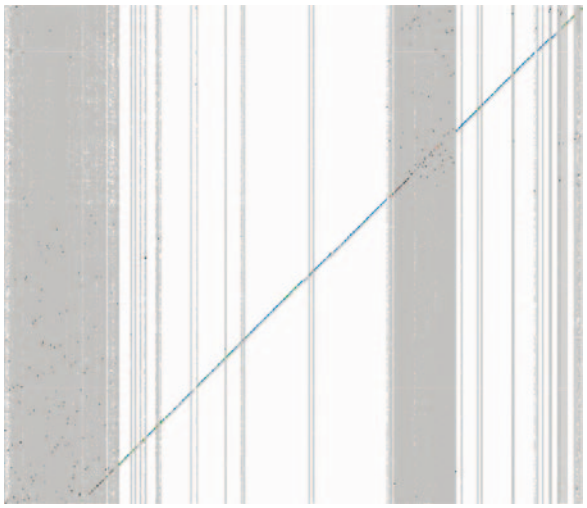

図 2：ブック照合に用いられたページ照合の分布．横軸の座標は寛政武鑑（寛政 7 ／1795）[200018828]，縦軸の座標は寛政武鑑（寛政 5 ／1793）[200018827]のページ番号（昇順）に対応し，灰色点に対応するペアはページ照合を行った．縦軸の武鑑には前欠があるため，横軸は途中から始まっている．

以上の画像照合機能を体験するデモ機能として，ファイル版とウェブ版の 2 つの機能が使える「差読のための画像照合サービス」を公開した [17]．ファイル版は PC に保存したローカル画像を 2 枚読み出して照合できる．またウェブ版は，URL でアクセスできる 2 枚の画像が照合できる．IIIF 形式の画像の場合，IIIF Curation Viewer から Image API の URL を取得することで，ウェブ版の照合が利用できる．しかし一般の IIIF ビューアの閲覧中に画像の URL が簡単にわからない場合は，画像を PC にいったんダウンロードしてから，ファイル版で照合することもできる．

5．ブック照合

5.1 ブック照合と同一板木判定

ブック照合の目的は，同一の板木から印刷されたページの組を判定し，その組に対してページ照合を行うことである．しかしどのページがどの板木から印刷されたかというメタデータは存在しない．また，版本が完全な形では残っておらず一部のページが欠けていたり，乱丁・落丁・増丁などでページの順番が入れ替わっていたり，ファイルナンバリングの問題でページ番号がずれていたりするなどの問題も頻繁に発生する．そのため，たとえ 2 点の版本が同じ武鑑であっても，同じページ番号が同じ板木に対応する保証はない．ゆえに同一の板木から印刷されたページのペアを自動的に判定するアルゴリズムが必要である．

最も単純なアルゴリズムは，すべてのページペアを総当たりで比較し，照合品質（正対応特徴点の数）が最大値となるページペアを選ぶというものである．しかしこのアルゴリズムでは，総当たりの場合の数が大きくなるという問題がある．例えば 200 丁（400 ページ）の版本の比較には 400 ×400＝16 万回のページ照合が必要となり，1 回のページ照合がたとえ 0.5 秒で完了したとしても，全体の計算には 1 日かかることになる．

そこでこの単純なアルゴリズムを 2 つの点から改良する．第一に，ページペアの探索範囲を絞り込むことで，ページ照合の回数を削減する．第二に，増丁がなければ同一板木のページペアは 1 つしかない，という制約を考慮したアルゴリズムを用いる．これらの改良について以下で述べる．

5.2 ページペアの探索範囲の絞り込み

同一板木のページペアは一般的には未知だが，実際には予測可能な場合が多々ある．例えば版本 A の m ページと版本 B の n ページが同一板木と判定できたとしよう．すると次のページペアは，版本 A の m+1 ページと版本 B の n+1 ページになることが予想できる．そこでこのページペアの正対応特徴点を調べ，その数が十分に大きければ，他のページペアは調べなくても近似解として正しいと考えられるだろう．このような予測が当たればページペアを総当たりで調べる必要がなくなるため，計算量を大幅に削減できる．

このアイデアを検証してみたのが図 2 である．これはブック照合において探索したページペアの分布を d3.js で可視化した図である．横軸と横軸が 2 点の武鑑のページ番号に対応しており，点はページ照合を行ったページペアを示す．斜めの線は少数のページ照合で同一板木のページペアを発見できた場合であり，探索範囲の絞り込みが効果を発揮している．一方，縦の帯は総当たりのページ照合が必要になった場合であり，同一板木のページペアが「存在しない」ことを証明するために総当たりのページ照合を行った場合に対応する．同一板木が存在しないというネガティブな結果は，最終的に無駄になるにも関わらず，ここに膨大な計算時間を要するというのは大きな問題である．この問題が未解決のため，現時点で絞り込み法が削減できる計算量は，全体の数割程度にとどまっている．

5.3 安定結婚問題

ページ照合ごとに照合品質を計算して RDB に登録すると，いよいよブック照合の目的である同一板木判定が実行できる．このタスクの要件は，できるだけ照合品質の高いページペアを選びつつ，同一板木のページペアは 1 個しか存在しないという制約を満たす，というものである．そこで注目したのが安定結婚問題である．

安定結婚問題とは，2 つのグループ（例えば男性と女性）間で，それぞれのメンバーの好みの順番が決まっている場合に，最適なマッチングを計算する問題である．これは経済学における資源割り当ての基本的な問題であることから，多数の研究が行われてきた．この問題設定をブック照合に当てはめてみると，版本 A と版本 B の 2 グループの間で，好みの順番にページペアを決定する問題となる．ここで好みの順番には，ページ照合で得られる正対応特徴点の数を用いればよい．このようにブック照合を安定結婚問題に帰着させると，古典的な解法である Gale-Shapley（GS）アルゴリズム[18]で高速に安定して解くことができる．

6．「武鑑全集」差読プラットフォーム

6.1 武鑑全集のための照合計算

「武鑑全集」に差読の機能を導入するため，「武鑑全集差読プラットフォーム」[19]を構築した．まずそのために行った照合計算を紹介する．武鑑全集の中から 336 点を選び，以下の条件を満たすブックペアに対して照合計算を行った．

1. 版型（横本か縦本か）が一致する．
2. 同一名称であり，国文研書誌 ID 順に並べた場合，前後 2 つ以内に入る．
3. 同一名称ではないが，国文研書誌 ID 順に並べた場合，前後に入る．

国文研書誌 ID は，おおむね出版年の順に付番された番号であることから，上記の方法は出版年が近い版本を対象に照合計算を行うことを意味する．この方法で選ばれたのが，666 件のブックペアである．

次にブック照合に用いる画像を準備した．縦本は見開き画像のため，半丁に対応するよう 2 枚の画像に分割した．一方，横本は半丁を撮影しているため分割はしない．またいずれの場合も，切り取り範囲を紙面サイズに対する固定割合で設定し，紙面の周囲を中心に切り取るようにした．その結果，照合用画像 143,616 枚のデータセットを構築した．その内訳は，111,114 枚の縦本画像（分割前は 55,557 枚）と 32,502 枚の横本画像である．

さらにこれらの画像に対して AKAZE 特徴検出器を適用し，検出した特徴点を RDB に保存した．得られた特徴点は全体で 67,071,993 個であり，これは画像平均で約 467 個の特徴点に相当する．ここまでがブック照合の準備作業である．

最後にブック照合を実行する．それぞれのブックペアに対してブック照合アルゴリズムを適用し，ページ照合ごとの照合品質と射影変換行列を RDB に登録する．そして全体の結果に GS アルゴリズムを適用し，同一板木判定を行ってブック照合を終了する．

アルゴリズムとしてはこのように完成したが，問題は計算量である．以上の計算を 666 件のブックペアに対して実行したところ，ページ照合の回数は 7100 万回以上に達し，その計算には 50 コア以上の計算サーバでも 10 日以上を要することがわかった．同一板木判定で残るページペアは最終的に 52 万件程度であるため，それよりも大幅に多い回数のページ照合を計算していることが問題の根源にある．安定結婚問題にはページ照合が双方向で必要となることを考慮しても，7100 万/52 万/2=68 倍も余分にページ照合を計算していることになる．

このように計算量が増大する原因は，5.2 節に述べたように，同一板木ページペアが「存在しない」ことの確認のためである．もし同一板木ページペアが存在しないことがあらかじめわかっていれば，これらをスキップして計算量を削減できる余地がある．一方，不完全な形で残る版本を照合するには，ページ番号が大きく飛ぶケースにも対応する必要があり，探索範囲をむやみに狭めることはできない．ページ照合の回数を減らすには，背景知識に基づき人間が探索範囲を制限するなど，人機分業の考え方も必要になる．

6.2 ブック照合とページ照合

ブック照合の結果を差読プラットフォーム[19]で利用する手順を紹介する．利用者はまずトップページにアクセスし，基準とする武鑑を選択する．次に選択した武鑑とすでにブック照合された武鑑のリストが表示されるため，ここから照合する武鑑を選ぶ．すると，ブック照合の結果として同一板木と判定されたページペアの一覧が表示される．個別のページペアには照合品質（正対応特徴点数）に応じた背景色を付与した．照合品質が高い順に，100 以上は青，50 以上は緑，20 以上は赤，それ以下は黒である．また同一板木ページペアが見つからなかった場合は灰，人間が手動で画像の位置合わせを改善した場合は紫である．

ページペアをクリックすると vdiff.js を用いたページ照合画面に遷移し，画像位置合わせの結果を表示できる．画像位置合わせの結果に満足できない場合は vdiff.js editor に遷移し，4 つの対応点を動かしながらプレビュー画像上で「ピント」が合う位置を探す．そして最適な対応点を決定して保存ボタンを押すと，4 つの対応点の情報が RDB に保存される．これはページ照合結果とは別のテーブルに保存され，ブック照合の更新とは切り離されて永続化されることになる．一方，照合品質が 20 未満の場合は照合に失敗している可能性が高いため，vdiff.js editor に直接遷移する．

6.3 ページ移動とブック移動

ページ照合画面では，ページ照合の移動先として，ページ移動とブック移動という 2 種類のリン

クを提供する．本論文の重要な貢献は，ページ移動とブック移動の両者に対して，同一板木という関係性を維持した遷移を実現した点にある．

まずページ移動の場合，片方の武鑑で前ページや次ページに遷移すると，遷移後のページでも同一板木が表示されるようにもう片方の武鑑も自動的にページを更新するため，同一板木を常に表示しながら版本をめくることができる．またブック移動の場合，片方の武鑑を固定したまま，もう片方の武鑑を切り替え，同一板木のページペアを表示する．この機能は資料横断的に同一板木を追跡するために不可欠なだけでなく，差分翻刻の実用性を大きく左右する機能でもある．

そこで，差読のアプリケーションとして当初から想定していた差分翻刻を試行する実験を行うこととした．この実験の主な目的は，差分翻刻の実現可能性の評価であるが，実際にシステムを使い倒してみる（ドッグフーディングする）ことで，システムの改良を進めることも狙っている．

6.3 差分翻刻の試行

差分翻刻の試行として，一つの大名家を選んでデータを通時的に作成した．具体的には尾張藩（徳川家）の当主名と居城地，参勤交代時期の項目を対象に，武鑑全集のすべての版に含まれるデータを時系列データとして翻刻した．スタートは本朝武鑑 （[貞享３]／1686）である．そこから国文研書誌 ID の昇順で武鑑を差読プラットフォームで閲覧し，尾張藩の情報に更新があればそれをエクセル表に書き込む．もしブック移動の機能がきちんと動作すれば，ブック移動に伴って尾張藩の情報を探し直す必要がなくなるため，翻刻作業が効率化することが期待できる．

差分翻刻の有効性を評価するには，ブック移動の正確性の評価が必要である．これは，提示されたブック移動が正しいか（precision），あるいはすべてのブック移動を拾えているか（recall）という観点から評価できる．しかしここで問題となるのが，ブック移動が表示されない場合である．これはブック移動を拾えなかった場合だけでなく，そもそも同一木版ページが存在しない場合も含んでいる．この両者を区別するには，木版の連続性という側面も評価する必要がある．

木版の連続性では，同一木版ページがブックに存在するかを評価する．版本書誌学の用語でいえば，修・印レベルの違いは同一木版と判定し，刊レベルの違いは同一木版ではないと判定する．そして本論文で用いた画像照合の手法は，原理的に刊の違いを越えた照合ができない．そのため，たとえブック移動ができなくても，アルゴリズムの責任ではないと言える．また，ページの欠損や落丁が原因で同一木版ページが存在しない場合もある．ブック移動の正確性の評価では，これらの場合を除外せねばならない．

こうした問題を詳細に検討した上での評価は今後の課題であるが，予備的な評価では，おおむね 6 割から 7 割程度の武鑑に対して正しいブック移動が表示できていた．また残りのほとんどは木版の連続性に問題がある場合だった．ゆえに precision と recall の観点では，ブック移動は有効な手法であると評価できる．

今後は差分翻刻の実利用を通してユーザインタフェースの改良を進めるとともに，より大規模な差分翻刻により時系列データを構築し，差分翻刻の有効性の評価を行いたいと考えている．

7．ブックバーコーディングに基づく版本照合研究基盤へ

最後に，本論文で提案した手法を，全く異なる視点からも意味付けしておきたい．我々は本論文の提案手法を「ブックバーコーディング法」と呼ぶ．これは生物多様性研究の分野で種の同定に使われる「DNA バーコーディング」[20]にヒントを得た命名である．DNA バーコードとは種ごとに特異な DNA 配列のことを指し，BOLD（Barcode of Life Data Systems）データベースにはすでに様々な種の配列が登録されている．ある DNA 配列の種を調べるとき，その配列を DNA バーコードのライブラリと比較することで，種を同定することができる．

このように，種に特異的な配列をデータベースに登録しておき，未知のデータが入ってきたときはデータベースと比較して種を同定するというのが DNA バーコーディングの枠組みである．これは，ページに特異的な特徴点配列をデータベースに登録しておき，未知の画像が入ってきたらデータベースと比較して板木を特定するという，本論文の枠組みと類似しているのではないだろうか．このアナロジーが，ブックバーコーディング法という名前の由来である．

しかしアナロジーはそれだけにとどまらない．生物種の進化の過程を系統樹というモデルで表現するのと同じく，版本の「進化」の過程も系統樹というモデルで表せるのではないか．そう考えると，未知の版本を系統樹上の適切な位置に対応させるという問題が，ブックバーコーディングのアナロジーで自然に理解できるようになる．

これは，生物情報学の考え方を人文情報学に導入する研究手法とも言える．筆者は以前に「スクリプトーム」という概念を提唱し[21]，データの網羅的な解析から新しい知識を得るというデータ駆動型研究において，生物情報学の長い歴史と多くの成果から学べることは多いと主張した．版本照合もそうした問題の一つと言える．

本論文で提案したブックバーコーディング法は，武鑑全集だけでなく一般の版本にも適用可能な汎用性を備えている．将来的には，武鑑全集差読プラットフォームを発展させた版本照合研究基盤を構築したいと考えている．

８．おわりに

本論文は，版本を対象としたページ照合とブック照合のアルゴリズムを提案し，武鑑全集の差分翻刻に適用して有効性を確かめた．またこの手法の全体をブックバーコーディング法と命名し，それを版本照合研究基盤へと拡大する展望を述べた．現状のブックバーコーディング法は計算量が多いという問題を抱えているため，この問題の解決を進めつつ，さらに大規模な差分翻刻を進めていく計画である．

謝辞

差分翻刻の試行には，合同会社 AMANE の寺尾承子氏の協力を受けた．本研究の一部には，科研費基盤研究(A)「歴史ビッグデータ研究基盤による過去世界のデータ駆動型復元と統合解析」（No. 19H01141）の支援を受けた．

参考文献

[1] 北本 朝展, 堀井 洋, 堀井 美里, 鈴木 親彦, 山本 和明, 時系列史料の人機分担構造化：古典籍『武鑑』を参照する江戸情報基盤の構築に向けて，人文科学とコンピュータシンポジウム じんもんこん 2017 論文集, pp. 273-280, 2017.
[2] 北本 朝展, 画像データの分析から 歴史を探る −「武鑑全集」における「差読」の可能性−, 歴史情報学の教科書, 後藤 真, 橋本 雄太 (編), pp. 53-58, 文学通信, ISBN 978-4-909658-12-8, 2019.
[3] 北本 朝展, 人物データの分析――江戸時代のデータブック「武鑑」の構造化と歴史ビッグデータ解析――, 電子情報通信学会誌, Vol. 102, No. 6, pp. 569-571, doi:10.20676/00000350, 2019.
[4] Kitamoto, A. and Yamamoto, K., High-throughput Collation Workflow for the Digital Critique of Old Japanese Books Using Computer Vision Techniques, Sixth Annual Conference of the Japanese Association for Digital Humanities (JADH2016), 2016.
[5] Leyh, T. and Kitamoto, A., Computer Vision-based Comparison of Woodblock-printed Books and its Application to Japanese Pre-modern Text, Bukan, Tenth Conference of Japanese Association for Digital Humanities (JADH2020), pp. 53-59, 2020.
[6] Kitamoto, A., Book Barcoding for Differential Reading -Application to Woodblock Printed Books in the Bukan Complete Collection-, Eleventh Conference of Japanese Association for Digital Humanities (JADH2021), pp. 22-27, 2021.
[7] 藤實 久美子, 江戸の武家名鑑 武鑑と出版競争,吉川弘文館, 2008.
[8] ROIS-DS 人文学オープンデータ共同利用センター, 武鑑全集, <http://codh.rois.ac.jp/bukan/>（参照 2021-11-01）.
[9] 寛政武鑑, 須原屋茂兵衛, 寛政 1 年 (1789), doi:10.20730/200018823
[10] 北本 朝展, 鈴木 親彦, 寺尾 承子, 堀井 美里, 堀井 洋, 地理的史料を対象とした歴史地名の構造化と統合に基づく江戸ビッグデータの構築, 人文科学とコンピュータシンポジウム じんもんこん 2020 論文集, pp. 171-178, 2020.
[11] 北本 朝展, 本間 淳, Saier T., IIIF Curation Platform：利用者主導の画像共有を支援するオープンな次世代 IIIF 基盤, 人文科学とコンピュータシンポジウム じんもんこん 2018 論文集, pp. 327-334, 2018.
[12 中野 三敏, 和本のすすめ――江戸を読み解くために, 岩波書店, 2011.
[13]中野 三敏, 書誌学談義 江戸の板本, 岩波書店, 2015.
[14] Alcantarilla, P. F., Nuevo, J., and Bartoli, A., Fast explicit diffusion for accelerated features in nonlinear scale spaces. Trans. Pattern Anal. Machine Intell, 34:7, 1281–1298, 2011.
[15] Fischler, M. A., and Bolles, R. C., Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24:6, 381–395, 1981.
[16] ROIS-DS 人文学オープンデータ共同利用センター, vdiff.js, <http://codh.rois.ac.jp/software/vdiffjs/>（参照 2021-11-01）.
[17] ROIS-DS 人文学オープンデータ共同利用センター, 差読のための画像照合サービス, <http://codh.rois.ac.jp/differential-reading/>（参照 2021-11-01）.
[18] Gale, D. and Shapley, L. S., College Admissions and the Stability of Marriage. The American Mathematical Monthly, 69:1, 9-15, 1962.
[19] 武鑑全集差読プラットフォーム, ROIS-DS 人文学オープンデータ共同利用センター, <http://codh.rois.ac.jp/bukan/diff/>（参照 2021-11-01）.
[20] Moritz, C. and Cicero, C., DNA Barcoding: Promise and Pitfalls. PLoS Biol 2:10, e354, 2004.
[21] 北本 朝展, 歴史的典籍の検索機能の高度化, そしてスクリプトーム解析に向けて, ふみ, No. 6, pp. 4-5, 2016.