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

北本 朝展（情報・システム研究機構 データサイエンス共同利用基盤施設 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）

堀井 洋，堀井 美里（合同会社 AMANE）

鈴木 親彦（情報・システム研究機構 データサイエンス共同利用基盤施設 人文学オープンデータ共同利用センター）

山本 和明（国文学研究資料館）

本論文は古典籍「武鑑」を対象として，大規模データを構造化するための全く新しいワークフローを提案する．まず「武鑑」を時間的に連続して変化する「時系列史料」という新しい種類の史料と捉え，そこから生み出される多数のバージョンをソフトウェア工学の観点から解釈し，これを板本書誌学の概念と対応させる．次にバージョン間の差分を検出する方法としてテキストベースと画像ベースのアプローチを比較し，「武鑑」では特に画像ベース差分検出が有効であることを示す．さらに差分検出と差分翻刻を合わせたアプローチを「差読」と呼び，そのためのワークフローを「人機分業」として構築することが「武鑑」の構造化の鍵を握ることを論じる．その最初の成果を「武鑑全集」として 2017 年 11 月に公開した．

Structuring Time-Series Historical Sources by Human-Machine Specialization: Toward the Construction of Edo Information Platform Referring to “Bukan”

Asanobu KITAMOTO (ROIS-DS-CODH / NII)
Hiroshi HORII, Misato HORII (AMANE LLC.)
Chikahiko SUZUKI (ROIS-DS-CODH)
Kazuaki YAMAMOTO (National Institute of Japanese Literature)

This paper proposes a new workflow for structuring large-scale data, such as Pre-modern Japanese text “Bukan.” First, we define “Bukan” as a new type of historical sources called “time-series sources” that changes continuously over time, and interpret many versions associated with “Bukan” from the viewpoint of software engineering and make a mapping of those versions to the concepts of bibliography of Japanese old printed books. We then compare text-based and image-based approaches to the detection of difference, and propose a new concept “differential reading” that combines both the detection of difference, and differential transcription, to realize a workflow based on human-machine specialization, which is a key toward structuring “Bukan” The first preliminary result was released as “Bukan Complete Collection” on November 2017.

１．まえがき

国文学研究資料館が中心になって進める「歴史的典籍 NW 事業」は，10 年間で 30 万点に及ぶ古典籍の大規模デジタル化と，オープンデータとしての公開を基礎とした国際的な共同研究が進んでいる．情報・システム研究機構 データサイエンス共同利用基盤施設 人文学オープンデータ共同利用センター（CODH）も古典籍のオープンデータとしての公開に協力[1]するとともに，古典籍データのコンテンツを検索するための「ディープアクセス」技術の実現を目指して研究を進めている．

ディープアクセスを実現するにはデータの構造化が第一歩となる．古典籍をデジタル化した後，その中身をテキスト化（翻刻）し，標準的なデータ形式で整理するなど，生のデータを人間や機械が活用しやすい形に加工する構造化は，人文情報学研究における基本的な作業となる．さらに，くずし字が読めない現代日本人に対しては，オープンデータとして公開するだけでなく，それを利用可能な形式に構造化することが重要である．そこで CODH は，古典籍データの構造化に向けた研究に取り組んできた．具体的には，「日本古典籍字形データセット」では古典籍の文字情報と座標を構造化し，「江戸料理レシピデータセット」では料理本『万宝料理秘密箱 卵百珍』の料理記述を調理可能なレシピに構造化した．こうした視点から見れば，本論文は「武鑑」という古典籍を対

象に，人物情報や組織情報，地理情報などを構造化する研究に相当する．

人文情報学で扱うデータの複雑さを考えれば，こうした作業の完全な自動化はまだ困難であり，当面は人手に頼って進めざるを得ない．とはいえデータが大規模化するにつれ，人手による構造化にも限界が生じている．例えば「武鑑」の場合，200 年間も発行された多数の版を，すべて人手で構造化することは作業量の面で不可能に近い．そこで本論文は，人間と機械がチームを組んで問題を解く「人機分業（人機分担）」のアプローチを活用したデータ構造化ワークフローを提案し，「武鑑」を対象とした大規模データ構造化に着手する．

本論文の構成は以下の通りである．まず 2 章では本研究の対象である「武鑑」について，その特徴と課題を紹介し，本論文の中心的な概念である「差読（differential reading）」を導入する．続いて 3 章は人機分業アプローチの概念を紹介し，機械による画像ベース差分検出と人間による差分翻刻のアイデアを述べる．4 章はこれまでに行ったデータ構造化の成果，5 章は将来的な江戸情報基盤への発展に向けた構想を紹介し，最後に 6 章で本研究の貢献をまとめる．

2．「武鑑」の特徴と課題

2.1 「武鑑」とは

「武鑑」の特徴とその価値については，『江戸の武家名鑑』[2]に詳しい．

> 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．

この記述は「武鑑」という出版物の特徴を端的にまとめている．

第一に，データベース的な性格を持つという点である．現代の『会社四季報』のように，これは大名家や幕府役人に関するデータを，項目ごとに構造化して随時更新するタイプの書籍である．ゆえに，個々の版から読み取れる情報だけではなく，版間の差分情報も人々の出世等を示唆する情報として活用することができる．

第二に，長期的に出版され続けたという点である．多数の武鑑の情報を蓄積すれば，江戸時代 200 年間の大名家や幕府役人に関する長期的な時系列データベースとしての分析に価値が生じる可能性がある．

第三に，ロングセラーブックだったという点である．「武鑑」は全国的に需要が高く，発行部数も多かったため，現代まで残存する可能性が高い．

第四に，改訂の頻度が高かったという点である．情報の更新や訂正が頻繁に行われたため，たとえ誤りが生じても修正しやすく，実世界に起きた変化を追跡しやすいことが期待できる．

ここで情報の信頼性についても検討しておきたい．「武鑑」は須原屋茂兵衛や出雲寺和泉掾といった版元から出版された書籍であるため，幕府の公式記録ほどの信頼性は認められないかもしれない．とはいえ，幕府や諸藩の武士達がコアユーザであったことを踏まえれば，内容はおおむね信用に足るものと評価できよう．

もし「武鑑」の多数のバージョンを網羅的に収集し構造化した時系列データベースを構築できたらどうなるだろうか．それは，江戸時代の大名家および幕府役人に関する中核的な情報を蓄積したデータベースとして，日本文化研究の多くの分野にインパクトを及ぼすことが期待できるだろう．これほど魅力的な特徴を備えた史料にもかかわらず，なぜ網羅的なデータベースが存在しないのか．その大きな理由は，多数のバージョンを扱うことの困難さにある．

2.2 「武鑑」とバージョン

「武鑑」は 200 年以上にわたって出版され続けたロングセラーであり，しかも多い時は年に数度から月に数度と高頻度で改訂されたため，少しずつ内容が異なる多数のバージョンが存在し，それを個別に一つずつ翻刻していくのは現実的ではない．そこで従来の「武鑑」に関する研究では，書誌学的な調査として「武鑑」の発刊時期を推定する研究や，系図など特定の項目に関するより詳細な調査や他の史料との比較研究などが行われてきた．しかしこれらはあくまで限定的な調査にとどまり，この方法の延長で「武鑑」の全貌がつかめるとは考えにくい．

そこで重要となるのが，「武鑑」の複数バージョン間の差分を活用するという考え方である．差分とは，あるバージョンと別のバージョンとの間に生じた差異のことを指し，ソフトウェア工学では広く使われている方法である．差分に着目するメリットは，主に変化検出と情報圧縮の 2 点にある．まず変化検出とは，2 つのバージョンを比較して変化した部分のみを抽出することを指す．また情報圧縮とは，そのようにして抽出された情報が結果的に全体よりもはるかに小さなサイズになることを指す．差分（パッチとも呼ぶ）を基準バージョンに当てると別のバージョンが完全に

復元できるため，差分は変化に関する重要な情報をすべて含むコンパクトな情報表現として，有用性が高いと言える．

一方，武鑑のバージョンの構造はどうなっているか，ここでは中野による板本書誌学[3,4]の定義を参考に整理してみよう．江戸時代の出版は，板木に文字や絵を彫って印刷するという木版印刷が主流だった．版（板）権という言葉に端的に表現されるように，板木は財産として位置づけられるものであり，たとえ修正が必要でも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．このような修正などによって生じた変異を板本書誌学では(1) 刊（板・版），(2) 印（刷・摺），(3) 修（補・訂）の 3 つのレベルで区別する．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．中野は「刊・印・修の追究がテクスト・クリティックに必須である」と述べている[3]．

このような板本書誌学の定義とソフトウェアのバージョンとを比較してみると，「刊」はメジャーバージョン，「修」はマイナーバージョンに対応すると考えられる．メジャーバージョンが変わる場合，ソフトウェアの構造自体が大きく変わるためバージョン間の差分に着目する価値は小さいが，マイナーバージョンが変わる場合はバグ修正などの細かい修正が中心となるため，差分に着目することで圧縮された情報から多くの情報を読み取ることができる．一方，「印」に直接対応する概念はないが，印によって生じる差異は木版の摩滅欠損などによる刷り上がりの差異に対応することから，強いて対応付けるならば，同一のソースコードから別々のコンパイルオプションで生成されたバイナリから生じる実行結果の差異に近いと言えるかもしれない．

また，ソフトウェアのようにバージョン番号が付与されておらず，バージョン管理システムのように修正履歴を遡及することもできないため，任意の「武鑑」がどのバージョンに対応するかを正確に知る方法は存在しない．「武鑑」の書誌としての刊年は後世に推定されたものであり，史料批判を経てはいるものの，より有力な証拠が出てくれば修正されうるものである．そこで必要となるのが，「武鑑」の各バージョンとの差分から前後関係などを精密に推定可能な，「武鑑」専用バージョン管理システムであろう．こうしたシステムを構築することが，「武鑑」の網羅的な解析の鍵を握ることになると考えられる．

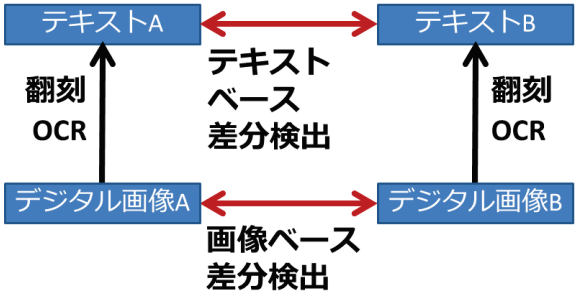

図 1　テキストベースと画像ベースの差分検出の比較．

2.3 差分検出

そこで次に「武鑑」の差分を検出する方法を考えてみよう．その候補となるのが，テキストベース差分検出と画像ベース差分検出である．図 1 はテキストベースと画像ベースの差分検出を比較した模式図であるが，テキストベース差分検出は翻刻されたテキストの比較から差分を検出する方法，画像ベース差分検出はデジタル画像の重ね合わせから差分を検出する方法である．

人文情報学における先行研究では主にテキストベース差分検出が用いられてきた．その理由は，写本が対象の場合，同一テキストが異なる書き手により異なるレイアウトで書かれるため，画像レベルでの比較は意味がなく，テキストレベルの比較しかできないからである．一方，「武鑑」のように連続的に板木が更新される版本では，前後のバージョンは基本的に同一の板木を利用し，修正も一部の置換にとどまるため，画像としての一致度が高く，画像ベース差分検出が有効に機能する．そこで本論文は，時間軸方向に連続的に更新される史料を「時系列史料」と命名し，そうした史料のための画像ベース差分検出技術を開発する．

画像ベース差分検出の利点は，非文字情報が使えるという点にもある．例えば中野は，コピーをつき合せることで版本の比較を行う際に，文字の点画の相違を比べるよりも匡郭や界線の欠損の有無を調べる方が簡便であるとして，非文字情報を用いた比較が有効なことを指摘している[3]．このような差異は，テキストベース差分検出では原理的に検出不可能である．

一方，テキストベース差分検出を大規模データに適用するには，古典籍 OCR の開発による全自動テキスト化の実現が望まれるが，文書単位の古典籍 OCR は未だに実現が困難な状態にある．第一に，古典籍で使われるくずし字は手書き文字のため印刷文字と比べて文字認識がはるかに困難，第二に，古典籍で使われる木版印刷はレイア

ウト設定が自由なため活版印刷と比べてレイアウト解析もはるかに困難であることから，現存する「くずし字 OCR」も指定した矩形内部に存在する文字（特に一文字）を認識するにとどまっている[5]．筆者らはくずし字認識のコンテストを開催するなどして，機械学習によるくずし字認識の研究を推進しているものの，文書単位の OCR の実現にはまだ時間を要すると考えている．それに対し，画像ベース差分検出は文字認識が不要でありレイアウトの複雑さの影響を受けにくいことから，時系列史料に対する差分検出には適した方法であると評価できる．

2.4 差読技術

このように，差分検出によって抽出された差分を読むという史料の読み方を，本論文では「差読（differential reading）」と呼ぶ．これは史料の読み方としては全く新しいアプローチであるが，時間的に連続変化する時系列史料の読み方としては理にかなったものである．

さらにここで強調したいのは，この読み方は機械の支援があって初めて実現できるという点である．ほぼ同一の画像に埋もれた微妙な差異を検出するのは多くの人間が苦手とする作業である．先入観が邪魔して差異を読み飛ばしたり，疲労によって注意力が散漫になったりするなど，苦痛を伴うこの種の作業を人間が長時間安定的に継続することは難しい．しかし機械は上記の問題とは無縁であり，むしろいくらでも続けられる得意な部類の作業と言ってよい．ゆえに，人間よりも機械の方が高品質な結果を生み出せる可能性さえあるのが，差読という読み方である．人文情報学では，人間による精読（close reading）に対して，機械による遠読（distant reading）という読み方がこれまで提案されてきたが，差読はこれらに加えた第三の読み方になりうると考える．

このような差読を実現するための技術を差読技術と呼び，その基盤となる画像ベース差分検出技術とそれを組み込んだワークフローを開発することが，本研究の主要な技術的課題である．

3．人機分業ワークフロー

3.1 人機分業とは

差読という新しい読み方では，人間と機械がそれぞれ得意とするタスクを分業しながら，時系列史料の大規模構造化を進めていくことになるが，これを本論文では「人機分業（または人機分担）」アプローチと呼ぶ．図 2 に示すように，史料全体を人間が翻刻する従来型ワークフローと比較し，人機分業に基づくワークフローは機械による差分検出の後に人間が差分を翻刻するという「差分翻刻」が基本となる．このワークフローでどのくらい作業量が削減できるだろうか．「武鑑」のバージョン間の差分に関する統計は，それ自体が今後の研究で明らかにすべき興味深い課題であるが，仮にバージョン間の差分を全体の 2%（50 バージョンで全体が入れ替わる）とすれば，1 バージョンごとの作業量は 50 分の 1 となる．このような圧倒的な効率化が実現できる点が，人機分業に基づく差分翻刻のインパクトである．

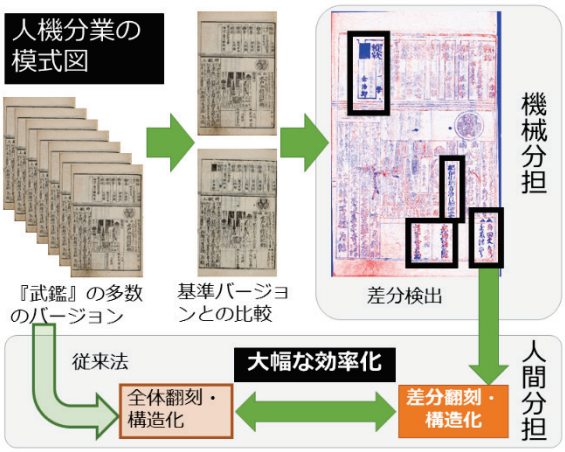

図 2　人機分業の全体像．差分検出を機械が分担し，翻刻・構造化を人間が担当することで，全体翻刻に比べ大幅な効率化となる．なお差分検出は予備実験の結果である．

人間と機械の協調というアイデア自体には既に多くの試みがある．例えばクラウドソーシングでは，多数の人間に同一の作業を分担してもらうことで，同時並行的に大規模作業を進めることができる．この枠組みでは，人間がすべてのタスクを実行し，機械は主にタスクの割り当てや完了タスクの品質検査などを分担することになる．一方，近年の機械学習では，人間が学習データを準備し，そこから学習したモデルによって機械がタスクを自動化するという分担もよく見られる．さらにウェアラブルシステムでは「人機一体」として，人間の意思を反映した機械が人間のタスクを助ける研究も多数行われている．

これら先行研究に対し，本論文で提案する人機分業とは，同一ワークフローの中で人間と機械の得意なタスクを見出し，それを細粒度に分割して分業を行う点に特徴がある．このアプローチでは，最適な分業ワークフローを個別タスクごとに考えねばならないが，人間と機械がそれぞれ得意とする作業が補完的であれば，分業が有効に機能することが期待できる．

3.2 機械：画像ベース差分検出

差読技術のうち機械側のコンポーネントは画

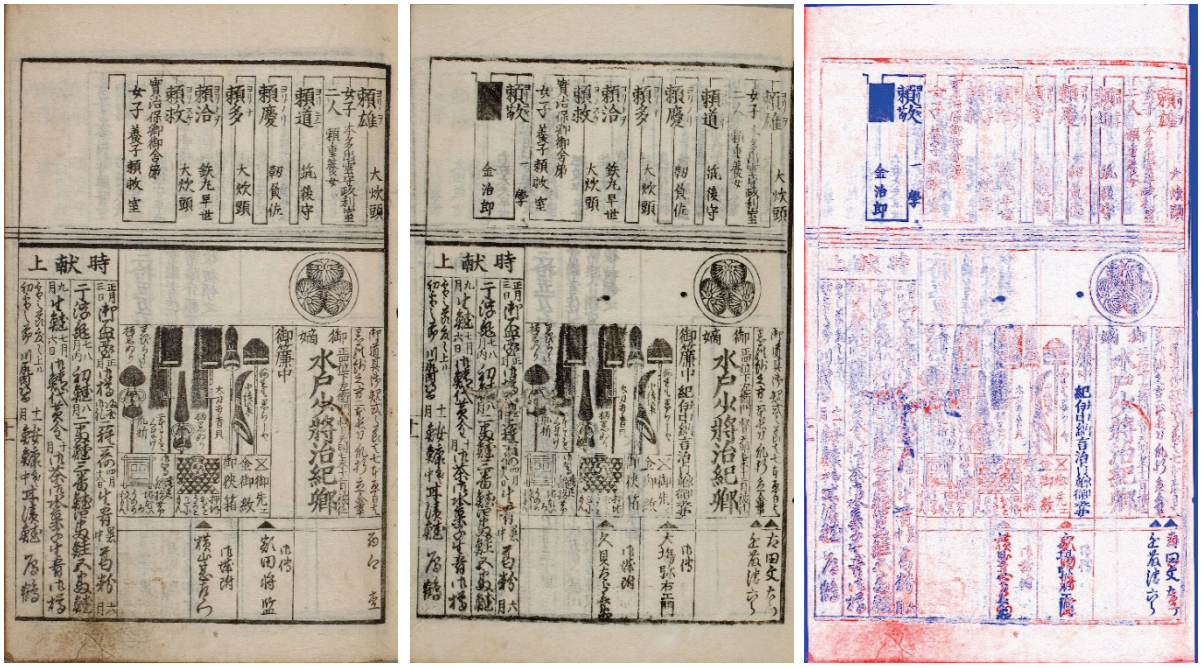

図 3 「武鑑」の比較例．左が『寛政武鑑』(1789)，中が『寛政武鑑』(1791)，右が両者を比較した結果．1789 年版のみ存在する部分は赤，1791 年版のみ存在する部分は青で着色している（ただし白黒印刷の場合は判別できない）．

像ベース差分検出となる．ここでは具体的に 2 つの課題を扱うことになる．

第一に，2 枚の画像を自動的に重ね合わせる技術の開発であり，これには特徴点の検出と記述，マッチングが鍵を握ることになるが，これらは画像処理の中核的アルゴリズムとして既に膨大な研究が行われているため，技術としてはほぼ確立していると言ってよい．またこうした史料の撮影では撮影条件がよく制御されているため，問題の難易度としても比較的低いといえるが，古典籍に特有の問題（木版の歪みや墨のかすれ，紙の劣化など）については，独自のロバストなアルゴリズムの検討は必要である．また改板の検出なども，時系列史料に特有の研究課題となる．

第二に，差分検出した結果を人間にどう見せるかという提示方法の問題がある．最も基本的な方法は，2 枚の画像の画素値の違いを強調する方法であり，例えば差分が大きい画素に有彩色（例えば青／赤），差分が小さい画素に無彩色（例えば白）を割り当て，差分を強調する方法が考えられる．またより進んだ方法として，差分が生じた画素の中からノイズを除去し，意味ある差分が生じた領域のみを自動抽出して，それを差分翻刻すべき領域に指定して翻刻プロセスに回すというワークフローも考えられる．もしここまで実現できれば，人間の意思決定負荷を下げて翻刻作業を高速化することができるだろう．ただし，領域抽出精度が低いと重要な差分を見落とす危険があるため，提示方法の違いがワークフロー全体の効率にどう影響するかなどを研究課題として考えていく必要がある．

3.3 人間：差分翻刻

差読技術のうち人間側のコンポーネントは差分翻刻となる．機械が差分を強調または抽出することで，翻刻すべき領域を大幅に削減できるところが差分翻刻の利点である．基準となるバージョンだけは全体翻刻が必要であるが，それに続くバージョンは差分だけを更新すればよく，改板を検出した場合のみ再度全体の翻刻をやり直すことになる．こうした改板がどの程度の頻度で行われたのかという統計は，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．

ここで注意を促したいのは，この方法が作業量の削減に寄与するだけでなく，品質の向上にも寄与しうるという点である．2 枚の画像を見比べて変化を検出することは人間にとって苦手な作業であるが，差分を強調すれば人間の注意を変化に引き付けることができ，意味のある差分を見落とすこともなくなる．つまりこのワークフローは，より高品質な翻刻に寄与できる可能性がある．

このような差分翻刻は現段階では実現していないが，将来的には IIIF (International Image Interoperability Framework) [6]を活用することで，画像の特定領域を指定した翻刻を可能とし

たい．そのウェブインタフェースは，複数バージョンを切り替え表示でき，差分を強調表示または抽出表示する機能を備え，領域に入力された構造化データをバージョン ID ごとにデータベースに反映できる，などの機能を備える必要がある．

4. 結果

4.1 画像ベース差分検出

2017 年 11 月現在，日本古典籍データセット [7]に収録されている「武鑑」は，須原屋茂兵衛が出版した『寛政武鑑』(1789, 1791, 1796, 1797, 1798, 1799)，『享和武鑑』(1804)，『文化武鑑』(1805)) の 8 点である．そこで本論文は，まず『寛政武鑑』(寛政 1 年, 1789) [8]と『寛政武鑑』(寛政 3 年, 1791) [9]の 2 点を選んで，その違いを比較することとした. 筆者らはすでに,「武鑑」以外の古典籍に画像ベース差分技術を適用した予備的な実験を報告[10]しているが，本論文は「武鑑」に対する予備的な実験の結果を報告する．

差分検出ソフトウェアの構築には OpenCV 2.4 を活用した．特徴点検出には FAST，特徴記述には BRIEF を利用し，マッチングにはハミング距離（全探索）を用いた．また 2 枚の画像の射影変換行列の推定には RANSAC を用いた．そして重ね合わせた画像に対して，1789 年版のみ存在する画素は赤，1791 年版のみ存在する画素は青で着色し，両方向の差分をカラーで強調するとともに，差分が小さい画素は白で表示することで背景化した．

その結果を図 3 に示す．この強調表示を見れば，1791 年版では左上の系図に追加があること，右下の人物名にも複数の追加や変更が存在することが一目瞭然である．この図を見ながら意味のある差分領域を矩形で囲み，それを差分翻刻プロセスに回せば，版面のごく一部の翻刻だけで情報を更新できることがわかる．

4.2 翻刻によるデータ構造化

本研究の人機分業では，長期的には差分翻刻を実現する計画であるが，現在はその前段階として，基準となる『寛政武鑑』(寛政 1 年, 1789) [8]の翻刻と構造化を進めている．ただし「武鑑」の全項目を翻刻することは困難なため，本論文では「武鑑」の「大名付」のみを対象とし，「役人付」は扱わないこととした．また「大名付」の掲載項目は年代によって変化するが，『寛政武鑑』については基本的な項目として，大名当主名，姓（本姓），領知高（単位:石），居城地，居城地-江戸間のルート・距離，参勤交代年月を選び，これらの項目のみを翻刻および構造化した．

さらにこの構造化データを現代の人々が活用

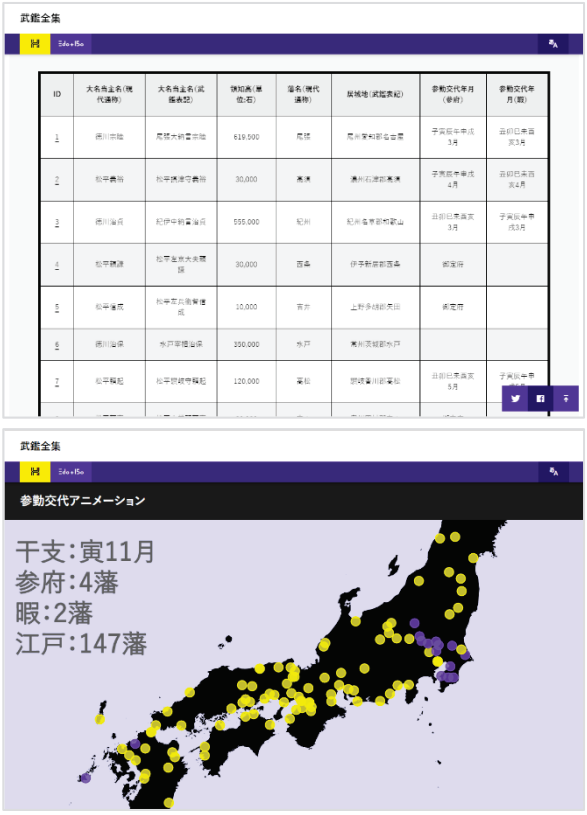

図 4　「武鑑全集」のウェブサイト．上が大名家一覧，下が参勤交代アニメーションのキャプチャ画像である．

しやすくするために，独自の項目を付け加えることとした．具体的には居城地の現在地に関する情報，現代の通称に関する情報などである．まず居城地に関する情報については，居城地が存在したと思われる場所を Google マップ上で探索し，その緯度経度と現在の住所を付加した．しかし記述が不十分などの理由から，17 の大名家については場所の特定に至らなかった．また現代通称については,『国史大辞典』・『角川新版日本史辞典』付録「近世大名配置表」等を参考に大名当主名および藩名の現代通称を付与した．

ここで重要な点として，大名家の ID の問題に触れておきたい．大名家の ID は各種の情報をリンクする典拠となる ID であることから，これを定めて共有することは江戸時代の情報を組織化する上で極めて大きな価値を持つ．しかし，大名家の粒度をどう定義するか，連続性をどう判定するかなど，歴史学上の問題をいくつも解決する必要が生じるため，こうした ID の定義には専門家集団の合議に基づく活動が必須であると考える．

そこで本論文の構造化では，典拠となりうるような ID の利用は想定せず,『寛政武鑑』の大名家の出現順に 1 から 264 までの大名家 ID を付与するという単純な方法を用いた．さらにこれを国

文学研究資料館が『寛政武鑑』に付与したDOIのサフィックスと組み合わせ，「武鑑」の数が今後増えても大名家を一意に特定できるIDとした．将来的に大名家IDが正式に標準化されれば，そこに特定バージョンごとのIDをリンクすることで，大名家に関する情報の時間的な変遷が追跡できるようにしたいと考えている．

5. 江戸情報基盤に向けて

5.1 「武鑑全集」の公開

こうして構造化したデータをウェブサイトに整理し，「武鑑全集」として2017年11月9日（大政奉還150周年の日）に一般公開した[11]．図4にウェブサイトの一部のページのキャプチャ画像を示す．現在は『寛政武鑑』（寛政1年, 1789）1点に関する情報しか掲載していないにもかかわらず，「武鑑全集」という壮大な名称を用いたのは，長期的には「武鑑」の時系列データベースに発展させたいとの意図が背景にある．

「武鑑全集」は，大名家に関する情報をリストまたは地図形式で一覧できるだけでなく，居城地の位置はGoogleマップと地理院地図でも確認できる．またデータベースを活用した可視化アプリケーションとして「参勤交代アニメーション」を公開した．これは「武鑑」の「大名付」に記される「参府」と「暇」の項目を参照し，大名家がどのような空間的・時間的パターンのもとで参勤交代していたかを可視化するものである．

「参府」と「暇」の項目には，どの干支の何月に江戸に到着し，江戸を出発するかが記されており，2年のうち1年は江戸に滞在するパターンに従う大名家が多いものの，1）江戸／大阪／京都に滞在しており参勤交代しない，2）毎年参勤交代する，3）3年に一度江戸に到着する，という大名家もある．特に江戸周辺の大名家は，正月前の12月に江戸に集合し，正月後の2月に解散する．こうした事実そのものは既知であるが，参勤交代アニメーションを眺めることで誰でも直感的かつ視覚的にこのようなパターンを把握できる点に，この可視化の価値がある．

また今回の構造化では活用できていないものの，武鑑に含まれる豊富な図版は今後の構造化においてぜひ活用したいと考えている．家紋や各種の印，模様などの図版は武鑑のもう一つの魅力であり，これらの利用は一般にも広がることだろう．

5.2 江戸情報基盤の構想

「武鑑」の魅力は江戸時代200年間の大名家や幕府役人に関する人物，組織，場所，時間などに関して，全貌とまでは言えないもののかなり網羅的にカバーしている点にある．そこに出現する人物などのエンティティに識別子を付与すれば，linked dataとして他の情報源を結合する基盤として使えると考えられる．例えば「武鑑全集」では『寛政武鑑』[8]を対象とした大名家IDを付与したが，同様の方法を拡大して様々なものにIDを付与していけば，江戸時代の人物や組織に関する概念ネットワークが構築できると考えられる．

一方，多くの情報を時間軸に紐付けられれば，「武鑑」が発行された200年間の任意の時期に移動して人物，組織，場所，時間などの情報を可視化する時系列インタフェース「江戸タイムマシン」が実現できると考えられる．例えば「武鑑全集」では参勤交代アニメーションを用いて，大名家の季節ごとの参勤交代パターンを可視化した．同様の方法を拡大して他の「武鑑」と比較すれば，参勤交代パターンが時代と共にどう変化したかなども，可視化できるようになると考えられる．

このように，「武鑑」から得られる情報を時空間に位置づけ，他の情報源とリンクする典拠とすることで，江戸時代の日本文化に関する多くの人文学研究分野で活用可能な，基盤的プラットフォームが構築できる可能性がある．本論文ではこれを「江戸情報基盤」と呼ぶ．

一方，こうした従来型の情報基盤にとどまらず，時系列史料を対象とした差読から得られる情報が，全く新しい発見につながる可能性も指摘しておきたい．例えば画素レベルでの版比較から，文字比較では見えなかった書誌学的な知見が得られるかもしれない．例えば藤實は，「武鑑」の摺り上がりが不均一な点から大名と板元との関係へと考察を深めた[12]が，同様に古典籍の内容に加えて出版物としての分析から，新たな知見を得るという方向性が有望である．例えば「武鑑」に間違いが存在することを逆手に取り，間違いがどの程度の速さで修正されたか，その速さに大名家ごとの違いがあるかなどを調査する．あるいは大名家や幕府役人の人事異動頻度がどのような長期変動パターンを示すのか，それには気候変動や疫病などが関係するかなどを調査する．こうした新しいリサーチクエスチョンに答えるための強力な調査ツールとして，翻刻を経由しない画像ベース差分検出が使える可能性がある．

このように「武鑑」の網羅的解析が新しい発見へのブレークスルーになる可能性を踏まえると，「武鑑」のような時系列史料はデジタル人文学時代に特有の重要史料ではないかと考えている．

5.3 オープンサイエンスの展開

こうした大きな計画を実現する鍵となるのが，オープンサイエンスの考え方である．具体的には，より多くの武鑑デジタルデータをオープンアク

セスとすること，そしてデータ構造化により多くの市民が参加できるシチズンサイエンスの環境を整備すること，この 2 点が課題となる．

「日本古典籍データセット」は現在 8 点の「武鑑」をオープンデータとして公開しているが，日本にはもっと多くの「武鑑」が各所に散在して残っている．こうした「武鑑」をネットワーク的に統合して利活用するには，標準技術としての IIIF が有用である．例えば各地の「武鑑」所蔵者がそれぞれデジタル化を行い，それをオープンなライセンスとともに IIIF で公開すれば，画像データを物理的に一か所に集めなくても，仮想的に集約しながら網羅的な解析が実現できる．

また「武鑑」というデータの大規模性を考えると，その構造化は一部の研究者だけで進められる分量ではなく，多くの人々の協力が必要となる．その際には，くずし字学習と組み合わせた翻刻や，地域活動と組み合わせた市民科学（シチズンサイエンス）など，単なる労働力の提供にとどまらない市民との出会いの場を作る必要がある．「武鑑」は日本の大部分をカバーする史料であり，人々の故郷にかつて存在した大名に関する情報を含むことから，構造化作業を「ふるさと翻刻」として地域ごとに分担するなど，地域の歴史を振り返りつつ，地域の歴史を全国に接続していく入口として，「武鑑全集」が使える可能性がある．

このようにオープンサイエンスの考え方を基本として，大規模構造化に多人数が協力できるようなプラットフォームを構築していくことが，「武鑑全集」を「江戸情報基盤」へと発展させていく鍵になると考えている．

6. あとがき

本論文は古典籍「武鑑」を対象として，大規模データを構造化するための全く新しいワークフローを提案した．まず「武鑑」を時間的に連続して変化する「時系列史料」という新しい種類の史料と捉え，そこから生み出される多数のバージョンをソフトウェア工学の観点から解釈し，これを板本書誌学の概念と対応させた．次にバージョン間の差分を検出する方法としてテキストベースと画像ベースのアプローチを比較し，「武鑑」では特に画像ベース差分検出が有効であることを示した．さらに差分検出と差分翻刻を合わせたアプローチを「差読」と呼び，そのためのワークフローを「人機分業」として構築することが「武鑑」の構造化の鍵を握ることを論じた．その最初の成果を「武鑑全集」として 2017 年 11 月 9 日に公開するとともに，それを将来的に江戸情報基盤に発展させていく道筋として，オープンサイエンスの考え方が重要であることを論じた．本論文で提案した新しいワークフローを活用し，大規模データの網羅的な解析がブレークスルーとなって，新しいリサーチクエスチョンから新たな発見が生まれることを期待したい．

謝辞

ノートルダム清心女子大学の藤實久美子教授には，「武鑑」に関する貴重なコメントを頂いた．本研究の一部には，科学研究費補助金 基盤研究 (B)「デジタル史料批判：エビデンスベース人文情報学のための連結指向型研究基盤,」(16H02920)，および国文学研究資料館研究開発系共同研究の支援を受けた．

参考文献

1) 北本 朝展，山本 和明：人文学データのオープン化を開拓する超学際的データプラットフォームの構築，人文科学とコンピュータシンポジウム じんもんこん 2016，pp. 117-124，2016.
2) 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.
3) 中野 三敏：書誌学談義 江戸の板本，岩波書店，2015.
4) 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.
5) 山本 純子，大澤 留次郎：古典籍翻刻の省力化：くずし字を含む新方式 OCR 技術の開発，情報管理，Vol. 58，No. 11，pp. 819-827，2015.
6) IIIF を用いた高品質／高精細の画像公開と利用事例：人文学オープンデータ共同利用センター，入手先〈http://codh.rois.ac.jp/iiif/〉（参照 2017-11-15）.
7) 日本古典籍データセット：人文学オープンデータ共同利用センター，入手先〈http://codh.rois.ac.jp/pmjt/〉（参照 2017-11-15）.
8) 寛政武鑑：須原屋茂兵衛，寛政 1 年 (1789)，doi:10.20730/200018823.
9) 寛政武鑑：須原屋茂兵衛，寛政 3 年 (1791)，doi:10.20730/200018825.
10) Asanobu KITAMOTO, Kazuaki YAMAMOTO: 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.
11) 武鑑全集：人文学オープンデータ共同利用センター，入手先〈http://codh.rois.ac.jp/bukan/〉（参照 2017-11-15）.
12) 藤實 久美子：大武鑑「大名付」と板元と大名家―江戸出版の仕組み―,徳川社会と日本の近代化，pp.335-363，思文閣出版，2015.