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――

Analysis of Person Data : Data Structuring and Historical Big Data Analysis of Bukan as Data Books of the Edo Period

北本朝展

1. はじめに

「武鑑」とは，江戸時代に出版された大名家及び幕府役人の名鑑である(1)．大名家や旗本家の当主・その家族・その家臣に関する人物情報だけでなく，大名家の石高，参勤交代時期，家紋や纏（まとい）のデザインなど，地理・経済・文化にわたる多種多様な情報を詰め込んだ，言わば江戸時代のデータブックである．「武鑑」は，17 世紀中頃に出版され始め，慶応 3 年（1867）の大政奉還までの 200 年以上の間，出版され続けた本である．「武鑑」の出版物としての特徴は，①データベース的な性格を持つこと，②長期的に出版され続けたこと，③ロングセラーブックだったこと，④改訂頻度が高かったこと，などの点にある．時系列的に長期間更新され続けた資料の網羅的な解析を通して，人物や組織などに関する細粒度のファクトデータを積み上げれば，経時的な統合分析から江戸社会の新たな側面が明らかになる可能性がある．

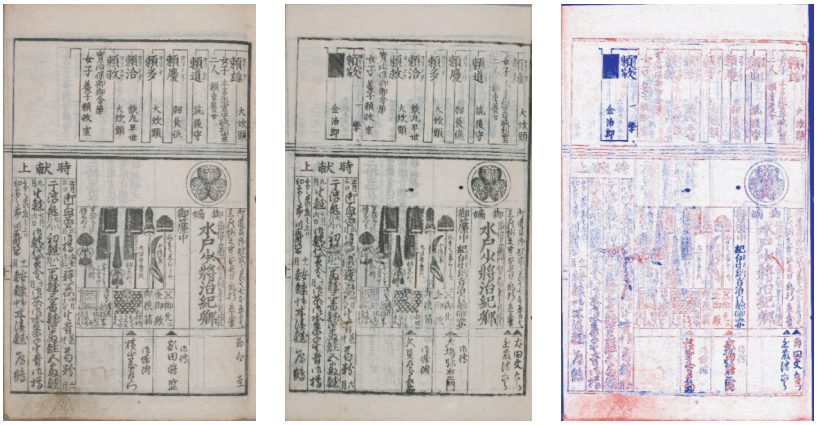
しかしそこに立ちはだかる大きな障壁が，「武鑑」のバージョンの問題である．「武鑑」は改訂頻度が年に数度から月に数度にまで増えた時期もあるため，その全てのバージョンを人間が一つずつ読んで分析することは困難だった．しかし「武鑑」のオープンデータ化が，この状況を変えつつある．まず，国文学研究資料館が中心になって進める「歴史的典籍 NW 事業」では，古典籍の大規模ディジタル化の一環として「武鑑」のディジタル化を進めている．そして 2018 年 12 月現在，ROIS-DS 人文学オープンデータ共同利用センター（CODH）の日本古典籍データセットでは，381 バージョンの「武鑑」を公開している．このオープンデータに対して人文情報学（ディジタルヒューマニティーズ）的なデータ駆動型解析手法を適用することにより，コンピュータが人間の「読み」を助ける新たな手法が生まれつつある．本稿はその背景にある「歴史ビッグデータ」の概念や研究開発課題を述べ，「武鑑」データの構造化に基づくアプリケーションを幾つか紹介する．

2. 歴史ビッグデータと構造化ワークフロー

「歴史ビッグデータ」(2)とは，「データを大規模に収集し，複数データを統合することで，世界を復元して解析する」という現代ビッグデータの考え方を過去に延長していく考え方を指す．現代ビッグデータと比べて歴史ビッグデータはデータの品質という面では多くの困難があり，現代ビッグデータの技術がそのまま適用できるわけではない．特に大きな課題が，歴史データをソフトウェアが処理しやすい形式に変換する「データ構造化」のワークフローである．

古典籍の場合，データ構造化の出発点はくずし字で書かれた古典籍を撮影した画像であり，これは人間がテキストを読む目的に適合した非構造化データである．これに翻刻や OCR（本小特集 2-1「文字データの分析―機械学習によるくずし字認識の可能性とそのインパクト―」を参照）を適用してテキスト化すれば，プレーンテキストという非構造化データが得られる．更にこれに対して人物や地名などをマークアップすれば半構造化データが得られ，これをスキーマに合わせて変換すれば構造化データが得られ，品質管理を行うことで解析に使えるデータ（analysis-ready data）がようやく得られる．ここまで到達できれば，後は現代ビッグデータの様々な手法が歴史データ分析に力を発揮できるだろう．つまりデータ構造化ワークフローの実現が，歴史ビッグデータ研究の鍵を握ることになる．

北本朝展　正員　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
E-mail kitamoto@nii.ac.jp
Asanobu KITAMOTO, Member (ROIS-DS Center for Open Data in the Humanities, Research Organization of Information and Systems, Tokyo, 101-8430 Japan).
電子情報通信学会誌　Vol.102 No.6 pp.569-571 2019 年 6 月



(a)『寛政武鑑』(1789)　(b)『寛政武鑑』(1791)　(c)(a)と(b)の比較結果

図 1「武鑑」の比較例　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．

同様の考え方は欧州においても大きな流れになりつつある．例えば EPFL（スイス連邦工科大学ローザンヌ校）では，「Big Data of the Past」をキーワードに Venice Time Machine プロジェクトが始まっている．これは，ベニス（イタリア）に残る 1,000 年以上の文書アーカイブの網羅的な分析を通して，時空間的に復元した都市を自由に移動可能なシステムを構築する野心的なプロジェクトである．このプロジェクトは欧州 33 か国，200 機関以上が参加する大規模なコンソーシアムに発展し，欧州の各地でローカルなタイムマシンの研究が始まっている．200 年に及ぶ「武鑑」を網羅的に解析する「武鑑全集」プロジェクト[3]も，同様の観点では「江戸タイムマシン」の一種とも言え，歴史情報基盤の新たなモデルを作り出すという目標には共通性がある．

3. バージョンと差分検出

3.1 木版印刷とバージョンの関係

ここで「武鑑」のバージョン問題を木版印刷の特徴と関係付けて紹介する．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，板木は財産としての価値が高く，再作成には多大なコストを要したため，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[4]では①刊（板・版），②印（刷・摺），③修（補・訂）の三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，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．

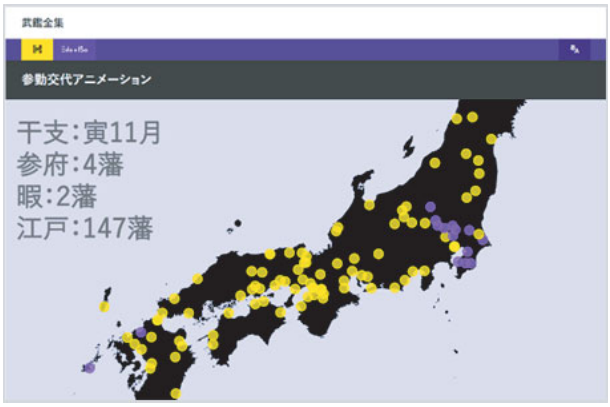
このような板本書誌学の定義と，一般的なソフトウェアのバージョンの定義を比較すると，「刊」はメジャーバージョン，「修」はマイナーバージョンに対応すると言える．一方，「印」は木版の摩滅欠損などによる刷り上がりの差異に対応するため，バージョンの対象とはならない．以上を踏まえると，最も重要なのが，「修」に相当するマイナーバージョン同士の比較である．メジャーバージョンの変更は大規模な構造変化を伴うため，バージョン間の差分に着目する価値は小さいが，マイナーバージョンの変更はバグ修正等の細かい修正が中心のため，差分が情報の圧縮表現として優れているからである．

3.2 画像ベース差分検出

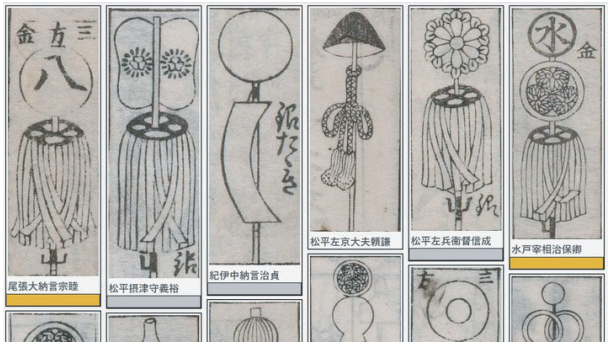
バージョン比較には一般的に差分検出が用いられる．ただし「武鑑」は分量が多いためテキスト化が困難であり，ソフトウェアで標準的なテキストベース差分検出は使えない．そこで，テキスト化が不要な画像ベース差分検出を用いる[5]．「武鑑」のように連続的に板木が更新される版本の場合，前後のバージョンは基本的に同一の板木を利用するため，板木を置換した部分のみを差分として強調できる．また匡郭や界線の欠損の有無など非文字情報の変化も，バージョンの前後関係を推定する重要なヒントとなる．

画像ベース差分検出は，技術的には 2 枚の画像を対応付ける問題に相当する．コンピュータビジョンの分野では，この問題に対する特徴点の検出と記述，マッチングに関する膨大な研究成果があるが，木版のひずみや墨のかすれ，紙の劣化，記録メディアの違いなど，古典籍の問題に対処するために頑健さを向上させる必要がある．

図 1 に『寛政武鑑』（寛政 1 年，1789）[6]と『寛政武鑑』（寛政 3 年，1791）の 2 点を選んで比較した例を示す．差分検出ソフトウェアには OpenCV を活用し，重ね合わせ画像に対して，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して背景化した．これにより，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．こうした



（a）　参勤交代アニメーション



（b）「纏」の大名家デザイン

前田家の江戸上屋敷
武鑑：上本郷五町目大手ヨリ三十二町
現在：東京都文京区本郷７丁目３－１
江戸切絵図：本郷湯島絵図

（c）　江戸上屋敷の『江戸切絵図』上へのマッピング

図２　「武鑑」の構造化データを用いたアプリケーション　参勤交代，大名家デザイン，古地図マッピングなど，江戸文化を多面的に探る切り口となる．

差分に基づき，変化が生じた部分だけをテキスト化すれば，データ構造化に必要なコストを削減できる．

4. 江戸情報基盤に向けて

「武鑑」の網羅的な分析に基づき江戸情報基盤を構築するというのが「武鑑全集」プロジェクトの目標である．例えば人物や大名家などに一意の ID を付与し，それらを時系列的にリンクすることで，どの人物がどのように出世したのかなど，様々な新しい問いに答えられるデータベースを構築できる可能性がある．ただし「武鑑」の記述を批判的に検証し，誰と誰が同一人物かを確定させるのは決して簡単ではない．歴史学者との共同作業によって地道に品質を向上させる長期的な研究が必要である．

また「武鑑」の構造化データからは，魅力的なアプリケーションも構築できる．図 2(a)のように大名家ごとの参勤交代時期を集めて可視化すれば，参勤交代の状況を時空間上で分析できる．図 2(b)のように大名家の家紋や纏などの「大名家デザイン」を一覧すれば，江戸の文化空間の特徴を浮世絵（錦絵）などと統合解析できる．更に図 2(c)のように大名家の江戸上屋敷の場所を当時の古地図である『江戸切絵図』(7)にマッピングすれば，江戸という都市を地理的に分析できる．これらのアプリケーションは，「武鑑」というデータブックから日本の文化を多面的に探る切り口を提供するものである．

5. お　わ　り　に

本稿は，「武鑑」という江戸時代のデータブックに対して，歴史ビッグデータという新しいアプローチを導入して江戸情報基盤を構築する目標を紹介した．そこで重要な役割を果たすデータ構造化については，研究者と機械と市民がそれぞれ得意なタスクを分業するための体制作りが重要な課題である．歴史ビッグデータの場合，機械による自動化はもちろん大切だが，データの背景にある専門知識に基づく批判的検証が必要な部分については，人文学と情報学との協働体制が不可欠である．

文　　献

（1）　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.
（2）　歴史ビッグデータ，http://codh.rois.ac.jp/historical-big-data/
（3）　武鑑全集，http://codh.rois.ac.jp/bukan/
（4）　中野三敏，書誌学談義　江戸の板本，岩波書店，2015.
（5）　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, Dec. 2017.
（6）　寛政武鑑，日本古典籍データセット，doi:10.20730/200018823, 1789.
（7）　“本郷湯島絵図，”〔江戸切絵図〕，国立国会図書館デジタルコレクション，doi:10.11501/1286676, 1849～1862.

（2018 年 12 月 31 日受付）

北本　朝展（正員）
本小特集 2-1（p. 568）を参照．