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 CODH／国立情報学研究所NII）

堀井 洋、堀井 美里（合同会社AMANE）

鈴木 親彦（CODH）

山本 和明（国文学研究資料館）

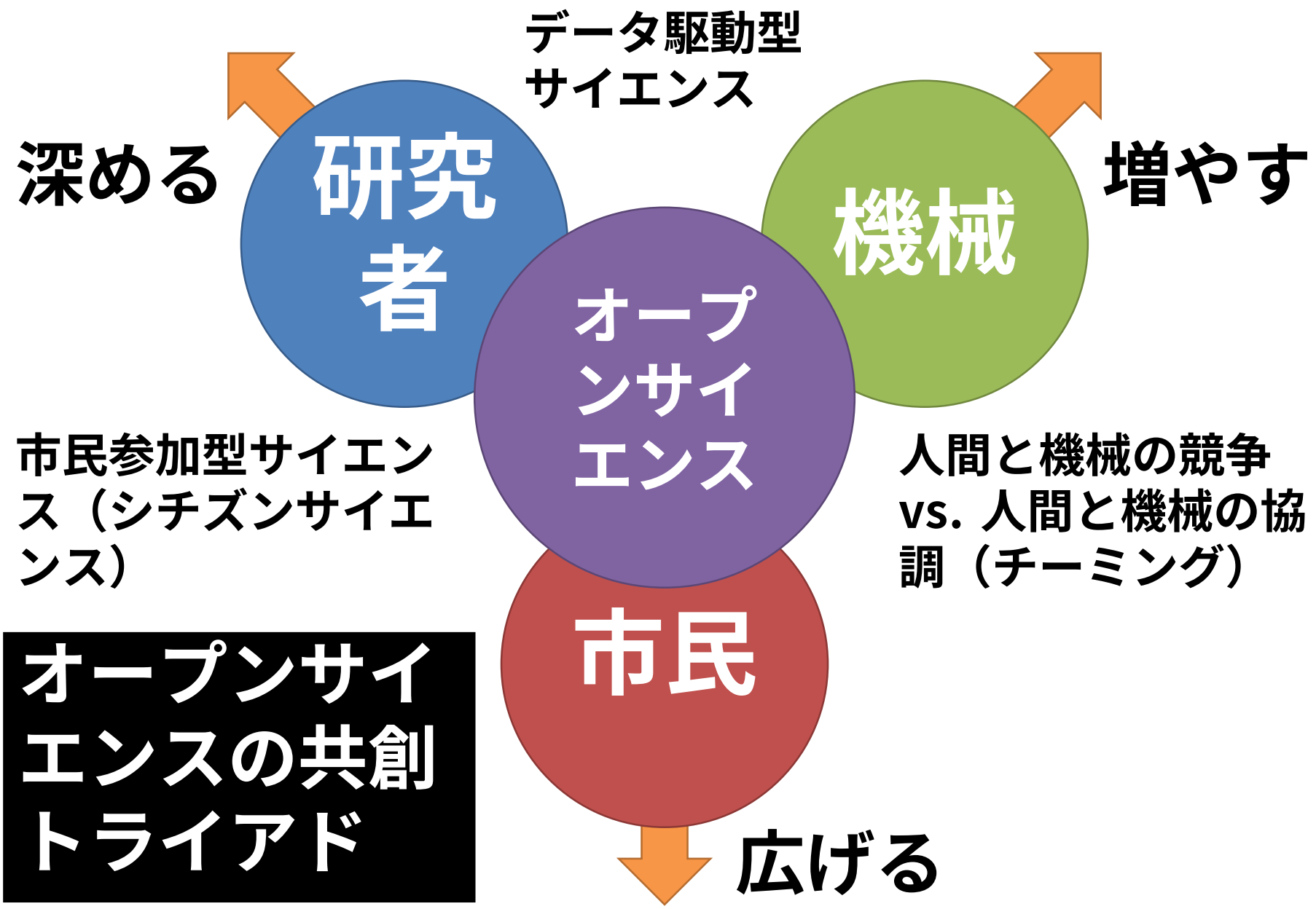
http://codh.rois.ac.jp/
Twitter: @rois_codh

人文学オープンデータ共同利用センター

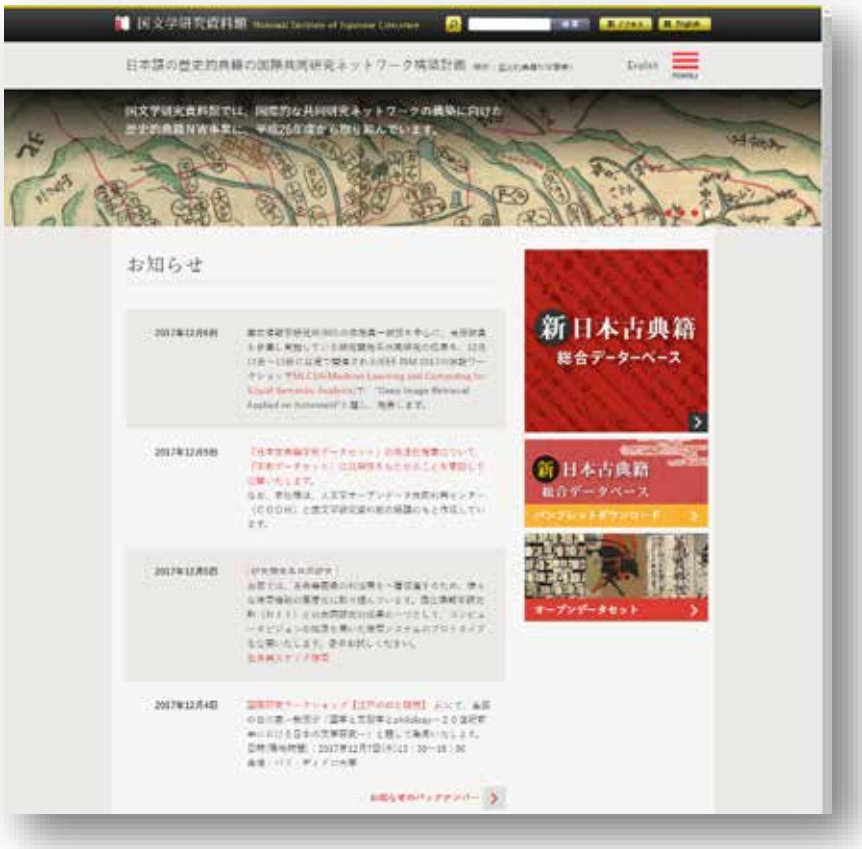
CODH http://codh.rois.ac.jp/

- 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内に、2017年4月1日に正式に発足。

1. **情報学・統計学の技術を用いて、人文学の研究を革新する。**
2. 人文学のデータを用いて、情報学・統計学の研究を革新する。


データ駆動型
サイエンス
深める
増やす
研究者
機械
オープンサイエンス
市民
広げる
市民参加型サイエンス（シチズンサイエンス）
人間と機械の競争 vs. 人間と機械の協調（チーミング）
オープンサイエンスの共創トライアド

NIJL-NWプロジェクト



http://www.nijl.ac.jp/pages/cijproject/

- **「日本語の歴史的典籍」約30万点を画像データ化。**
- **日本文化に関する、初の（？）ビッグデータ公開事例。**
- **CODH／NIIは、オープンデータ公開やデータの利活用のテーマで、国文研と協働。**

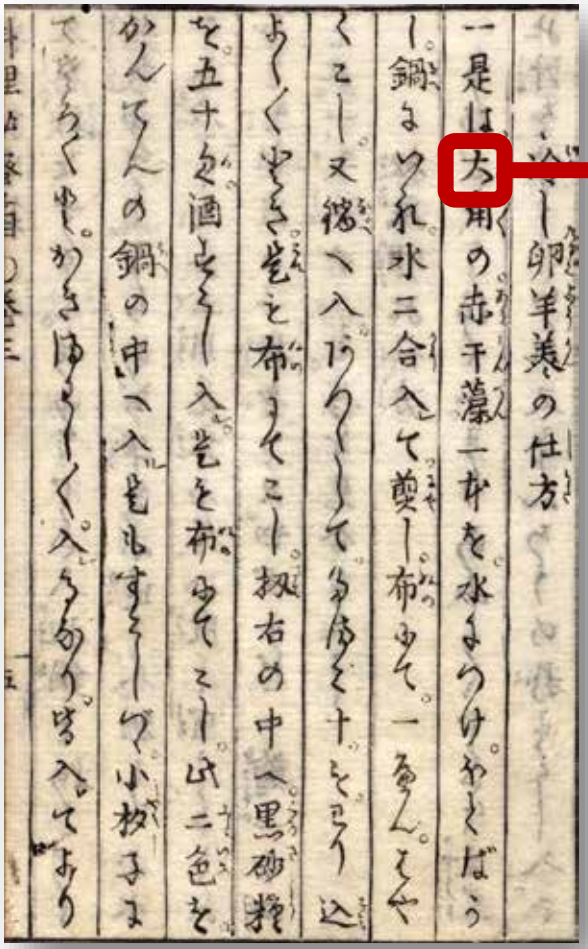
ディープアクセスとデータの構造化

- **ディープアクセス**：古典籍のメタデータに加え、内容分析を用いて直接アクセス。

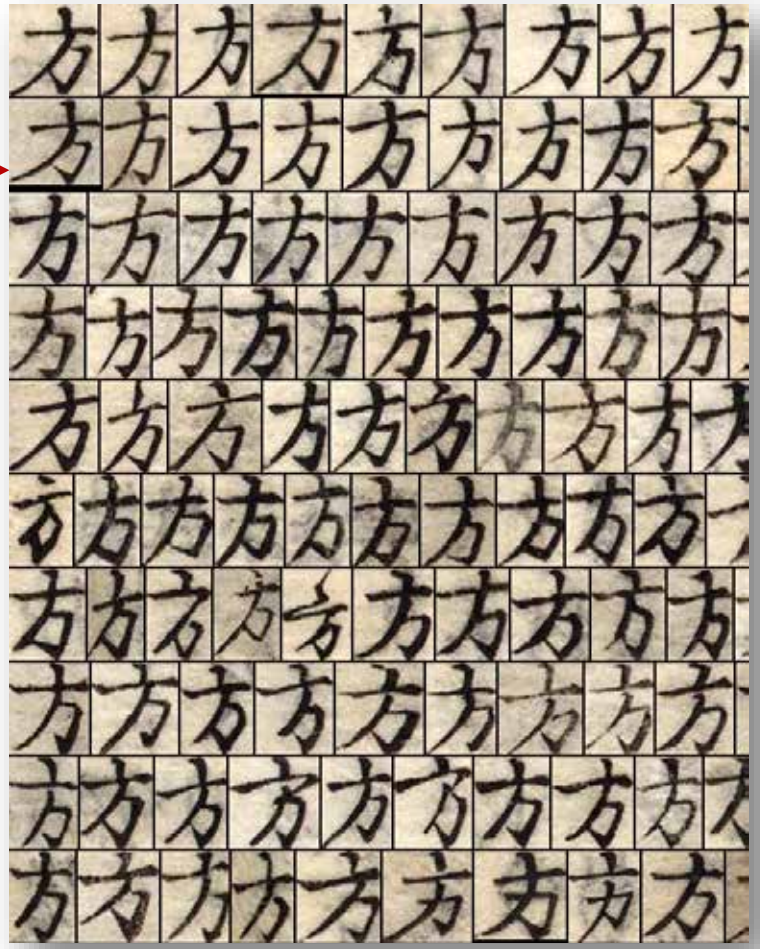
1. **日本古典籍字形データセット**：古典籍の文字情報と座標を構造化。
2. **江戸料理レシピデータセット**：料理本の記述を調理可能なレシピに構造化。
3. **武鑑全集（本研究）**：人物情報や組織情報、地理情報などを構造化。

機械のためのオープンデータ

http://codh.rois.ac.jp/char-shape/



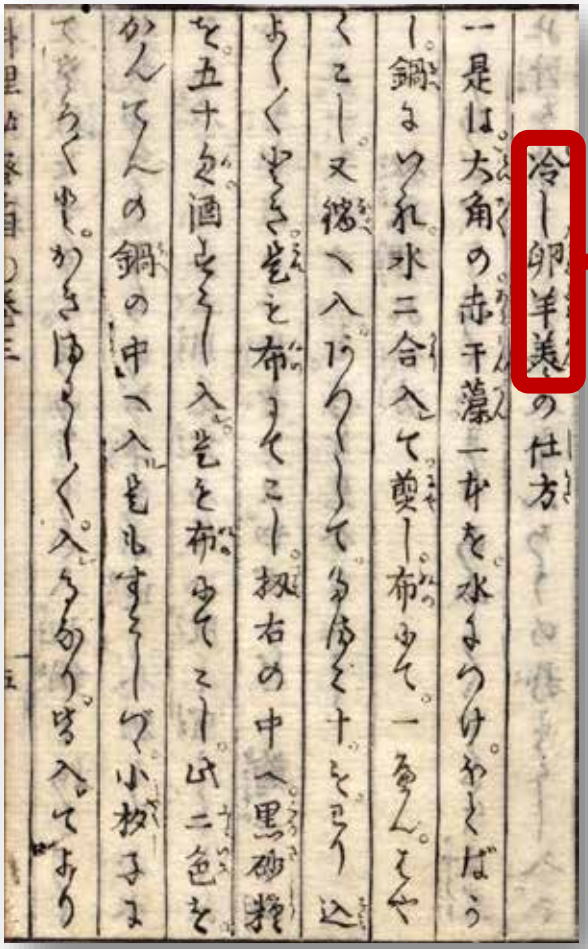
日本古典籍データセット
（国文研所蔵）



日本古典籍字形データセット
（国文研所蔵・CODH加工）

市民のためのオープンデータ

http://codh.rois.ac.jp/edo-cooking/



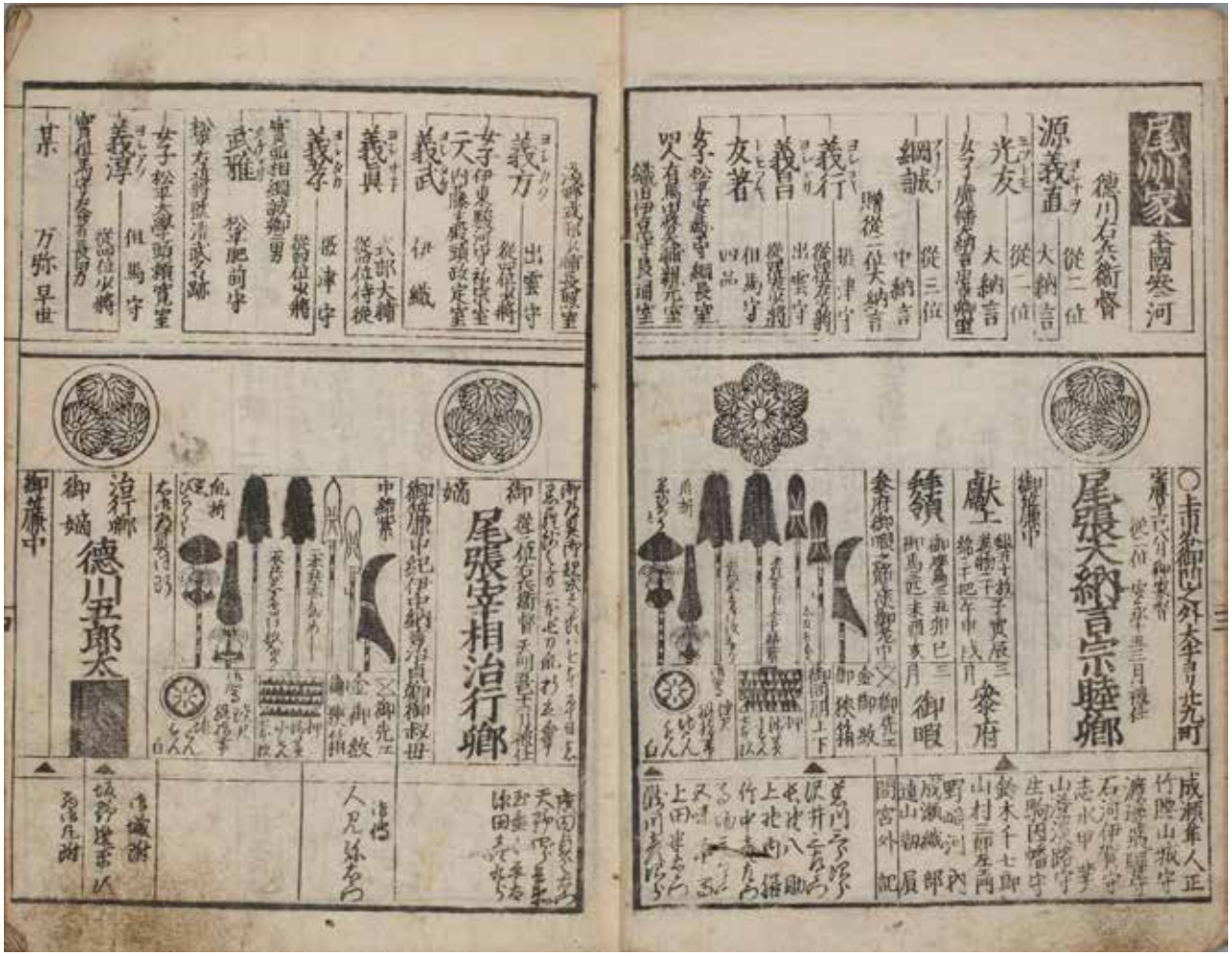
日本古典籍データセット
（国文研所蔵）

江戸料理レシピデータセット
（CODH制作）
日本古典籍データセット
（国文研所蔵）を翻案

「武鑑」のバージョンと「差読」

寛政武鑑（1789）

http://codh.rois.ac.jp/pmjt/book/200018823/



日本古典籍データセット（国文研所蔵）

「武鑑」とは？

1. 「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**。
2. 17世紀中ごろ～1867年の大政奉還まで、**200年以上の間出版され続けた**。
3. 実用書であり、**ロングセラーブック**。
4. 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。

参考：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

武鑑の価値と課題

1. もし「武鑑」を網羅的に構造化した時系列データベースを構築できたら？
2. **江戸時代の大名家・幕府役人に関する中核的データベース**になるのでは？
3. 日本文化の研究に役立ちそうなのに、なぜ網羅的なDBが存在しないのか？

- **→多数のバージョンが存在するため、そのすべてを構造化することは困難！**

古典籍とバージョン

- 江戸時代の出版は、板木に文字や絵を彫って印刷する**木版印刷**が主流。

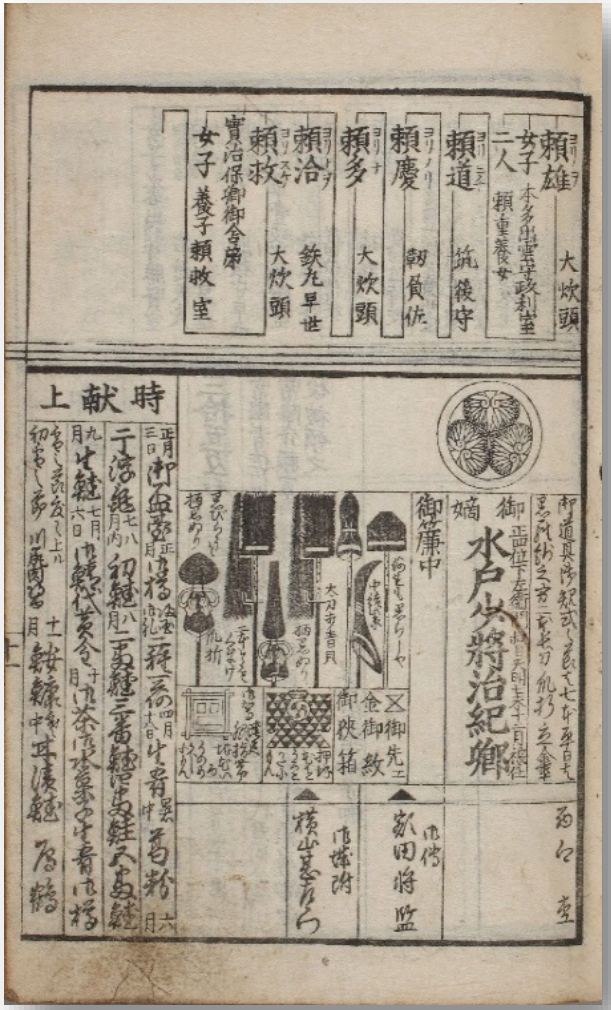
1. **刊（板・版）**：新しい板木を彫って本を刊行すること（**メジャーバージョン**）。
2. **印（刷・摺）**：既存の板木を使って本を刷ること（異なるバイナリ？）。
3. **修（補・訂）**：既存の板木に対して埋木などを使って部分的な修正を加えること（**マイナーバージョン**）。

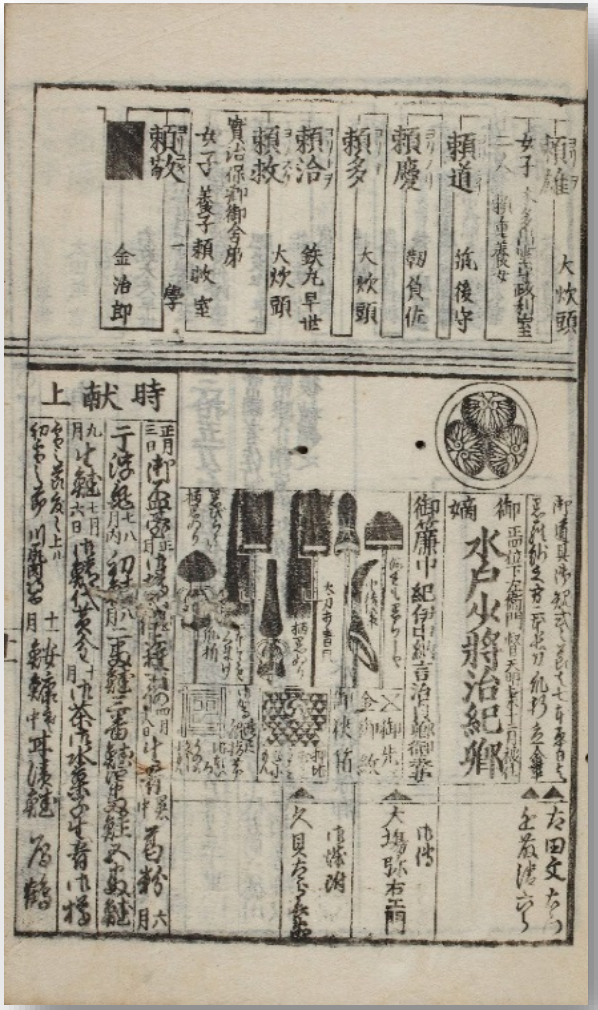
参考：中野 三敏：書誌学談義 江戸の板本, 岩波書店, 2015

武鑑とバージョン管理

1. 「武鑑」の刊年は後世の推定。同一年の複数バージョンは区別できない。
2. 任意の「武鑑」がどのバージョンに対応するかを知る手段はない。
3. 時間とともに更新される時系列史料のバージョン管理システムが必要。
4. バージョン管理システムの特徴は、差分による変化検出と情報圧縮にある。

異なるバージョンの比較



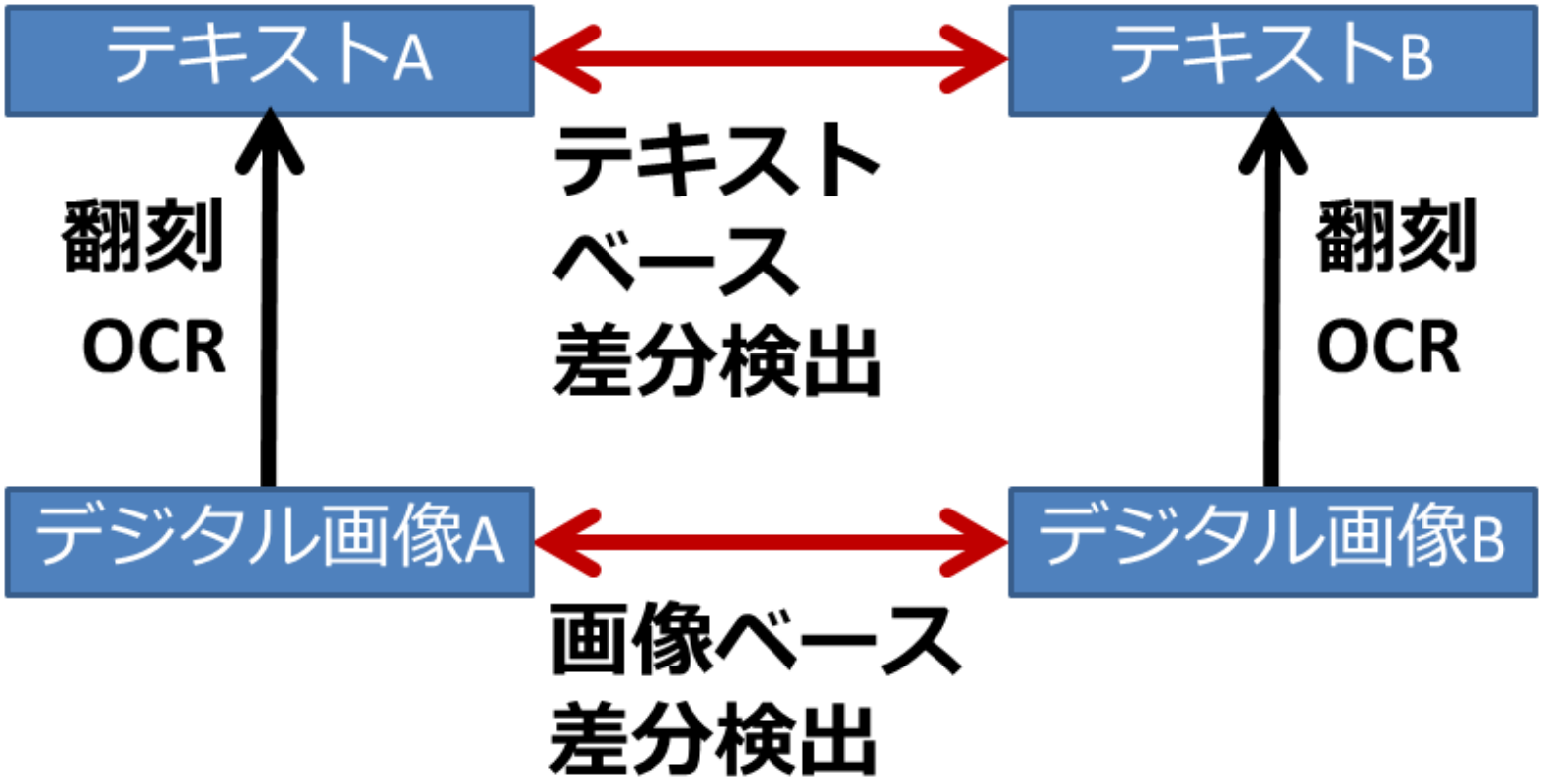


左が『寛政武鑑』（1789）、右が『寛政武鑑』（1791）。日本古典籍データセット（国文研所蔵）

差分検出の方法

1. **ソフトウェア工学**：diffなどの差分出力ツールや、gitなどのバージョン管理システムが広く使われている。
2. **テキストベース差分検出**：Versioning machine（TEI）など、差分に焦点を合わせたシステムは多く存在する。
3. **画像ベース差分検出**：写本の場合は適用が困難だが、版本の場合は同一板木を使いまわすため有効に働く。

テキストベースと画像ベース



画像ベース差分検出

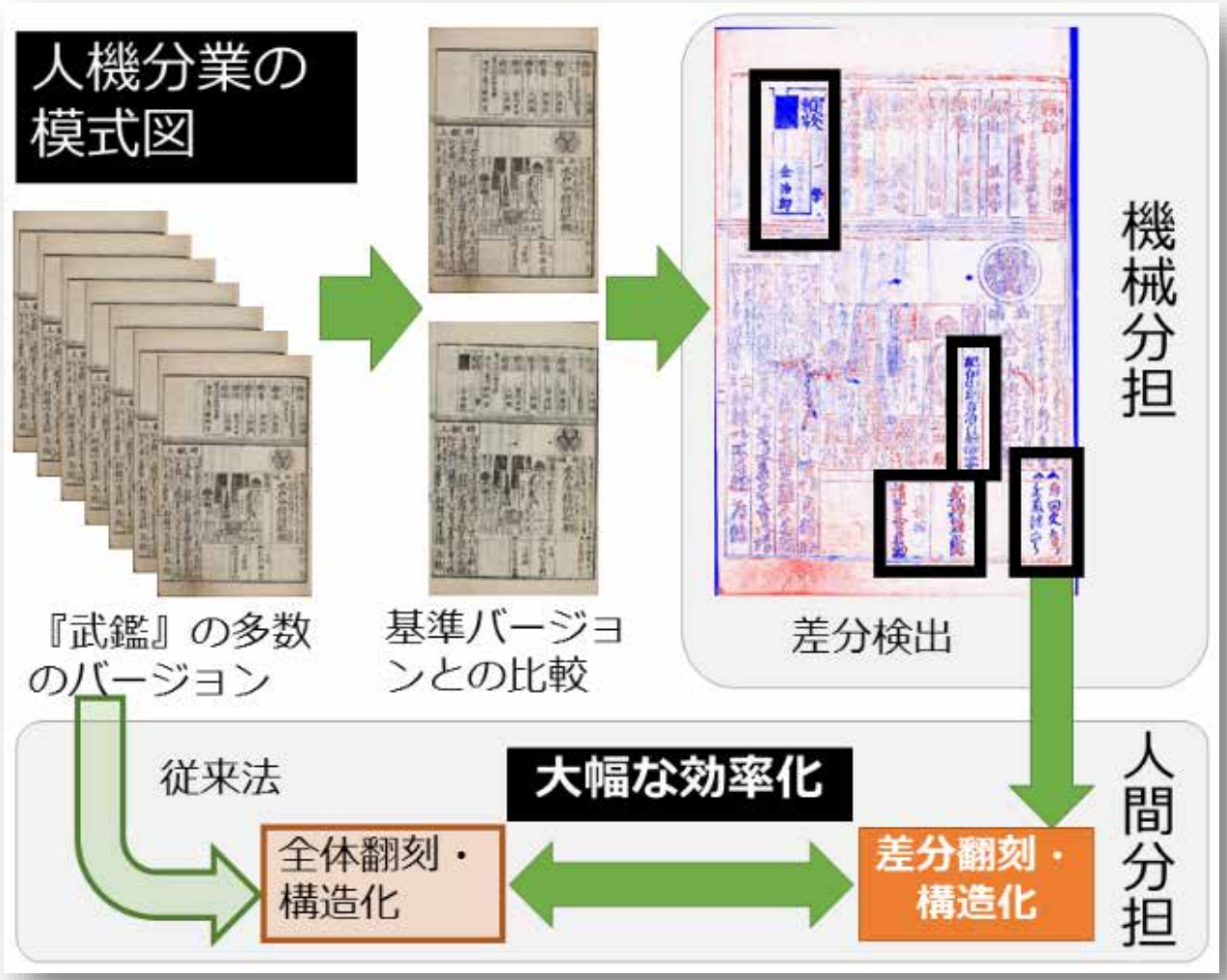
1. **テキスト化不要**：低精度なOCRや高コストな翻刻を回避可能。
2. **非文字情報の活用**：版本の比較では、文字の相違を比べるより、匡郭や界線の欠損の有無を調べる方が簡便。
3. **コンピュータビジョン**：特徴点を検出して対応を計算することは基本的問題。
4. **時系列史料に適した差分検出方法。**

差読（differential reading）

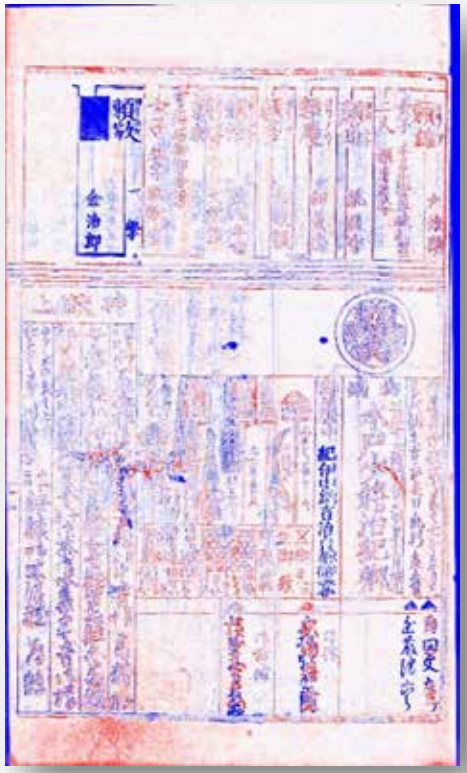
1. **遠読（distant reading）と同様、機械の支援を受けた新しい「読み方」の提案。**
2. **ほぼ同一の画像に埋もれた微妙な差異を強調し、人間が気づけるようにする。**
3. **人間が苦手、機械が得意とするタスク。人間よりも機械の方が、高品質な結果を生み出せる可能性が高い。**
4. **時系列史料に適した新しい読み方。**

人機分業ワークフロー

人機分業ワークフロー

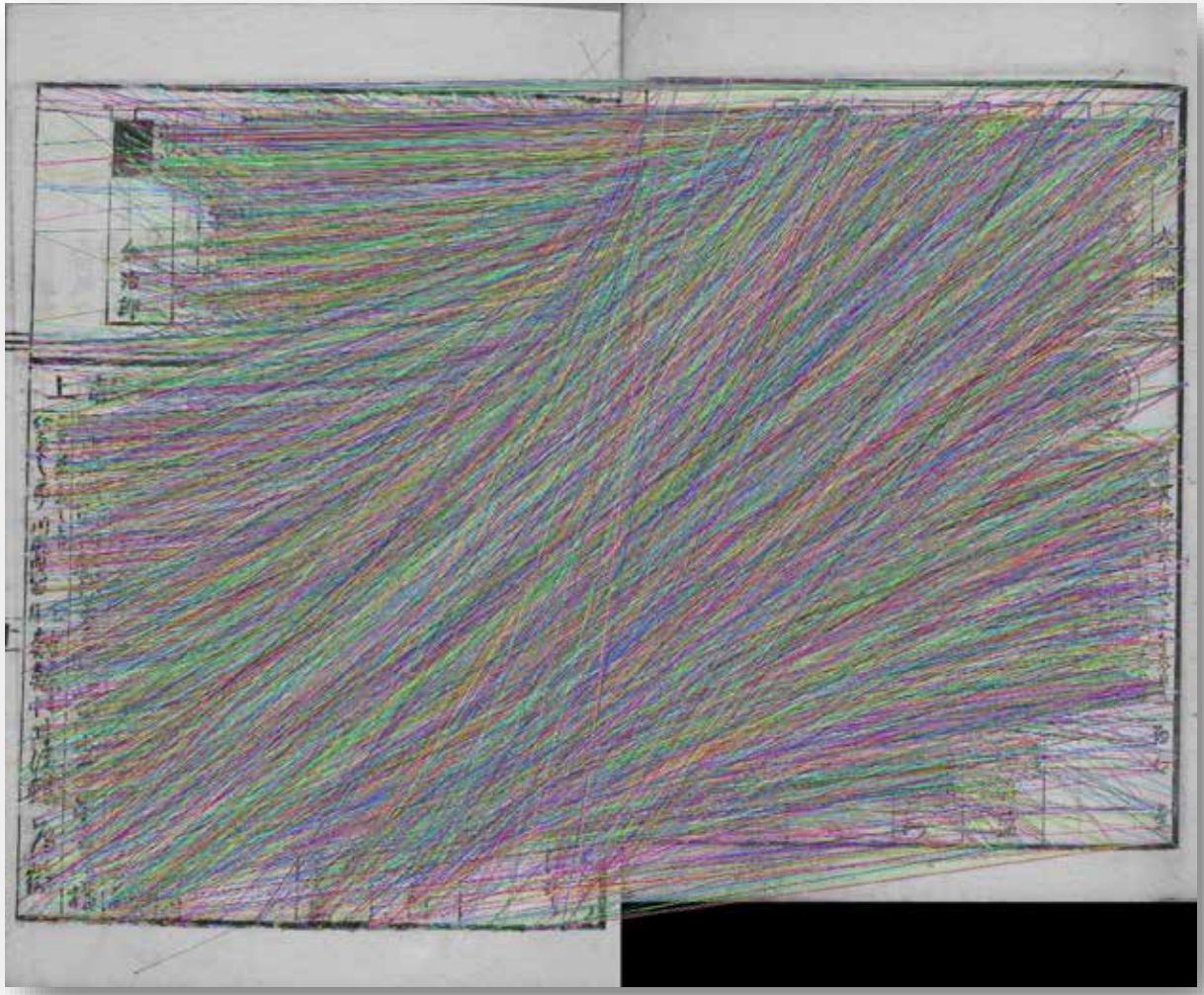

人機分業の模式図
『武鑑』の多数のバージョン
基準バージョンとの比較
差分検出
機械分担
従来法
大幅な効率化
全体翻刻・構造化
差分翻刻・構造化
人間分担

画像ベース差分検出



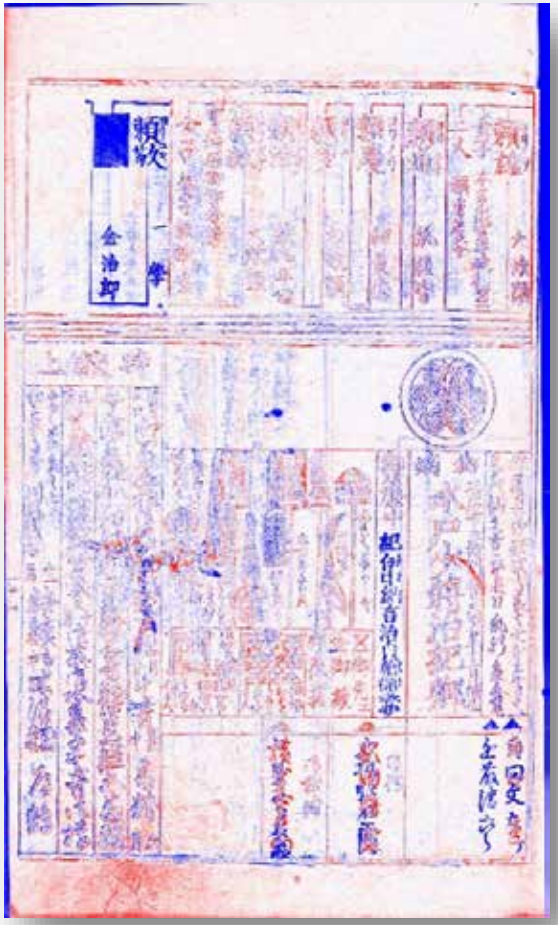
左が『寛政武鑑』（1789）、中が『寛政武鑑』（1791）、右が比較。1789＝赤、1791＝青。

画像の重ね合わせ



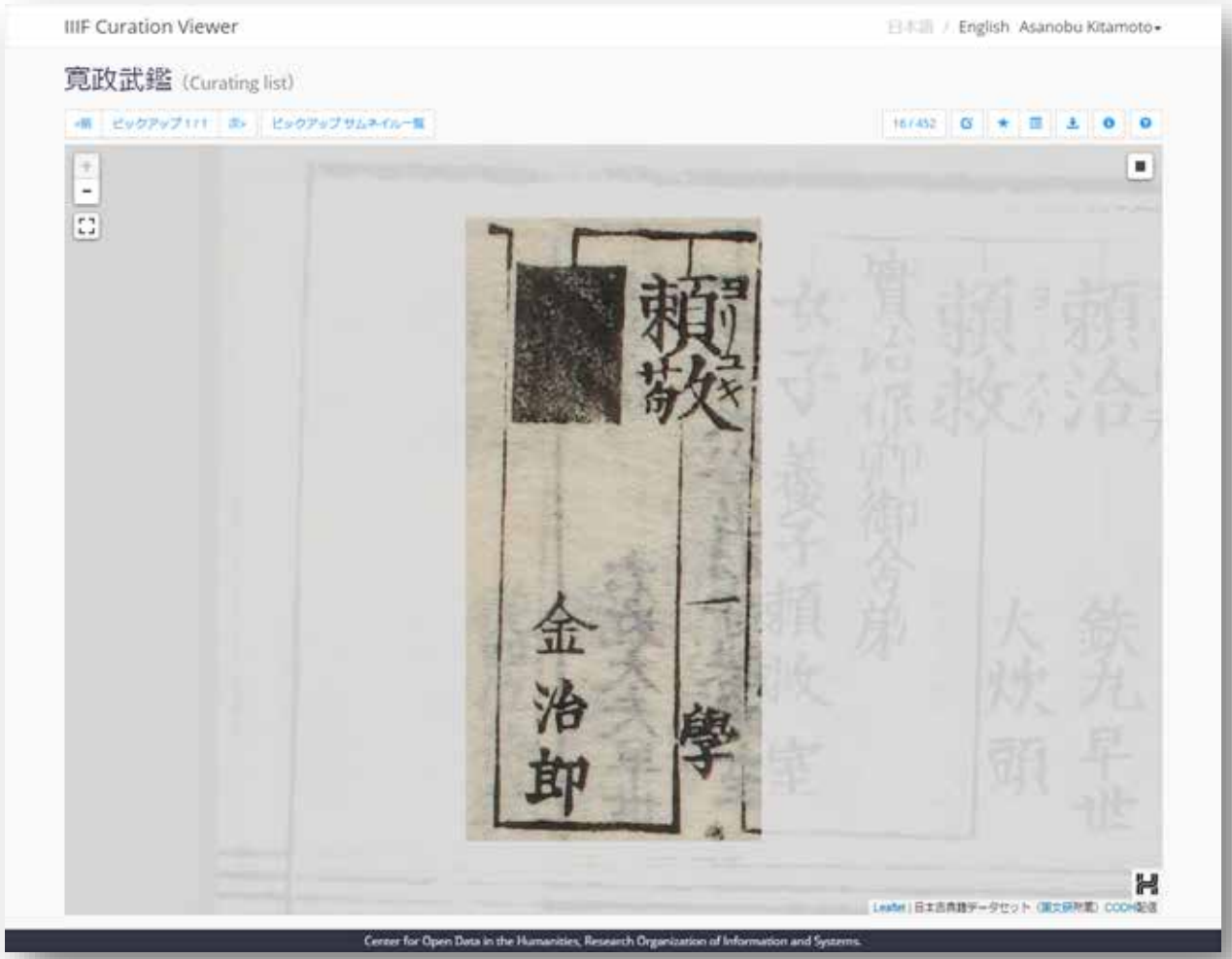
- OpenCV 2.4を利用。
- **特徴検出**：FAST
- **特徴記述**：BRIEF
- **マッチング**：ハミング距離（全探索）
- **射影変換行列推定**：RANSAC
- **差分表示**：青・赤でカラー強調。

差分翻刻



- **基準翻刻**：基準となるバージョンは、全部を翻刻する。
- **差分翻刻**：その後のバージョンは、差分のみを翻刻する。
- **改版検出**：板木の更新を検出したら、全部を翻刻しなおす。
- **メリット**：全体の2%が更新される場合、翻刻すべき量を1/50に削減できる。

差分翻刻システム構想



IIIF Curation Viewerを用いて、差分検出結果から**変化領域を自動生成**し、**差分翻刻を依頼**する。

武鑑全集と江戸情報基盤に向けて

Edo+150・武鑑全集

http://codh.rois.ac.jp/edo+150/

2017年11月9日公開
大政奉還150年記念

翻刻によるデータ構造化

- **基本項目**：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。
- **居城地の現在地**：Googleマップで探索。
- **現代の通称**：『国史大辞典』・『角川新版日本史辞典』など。
- **大名家ID**：標準がないため出現順に付与。本のDOIと組でユニークなIDとする。

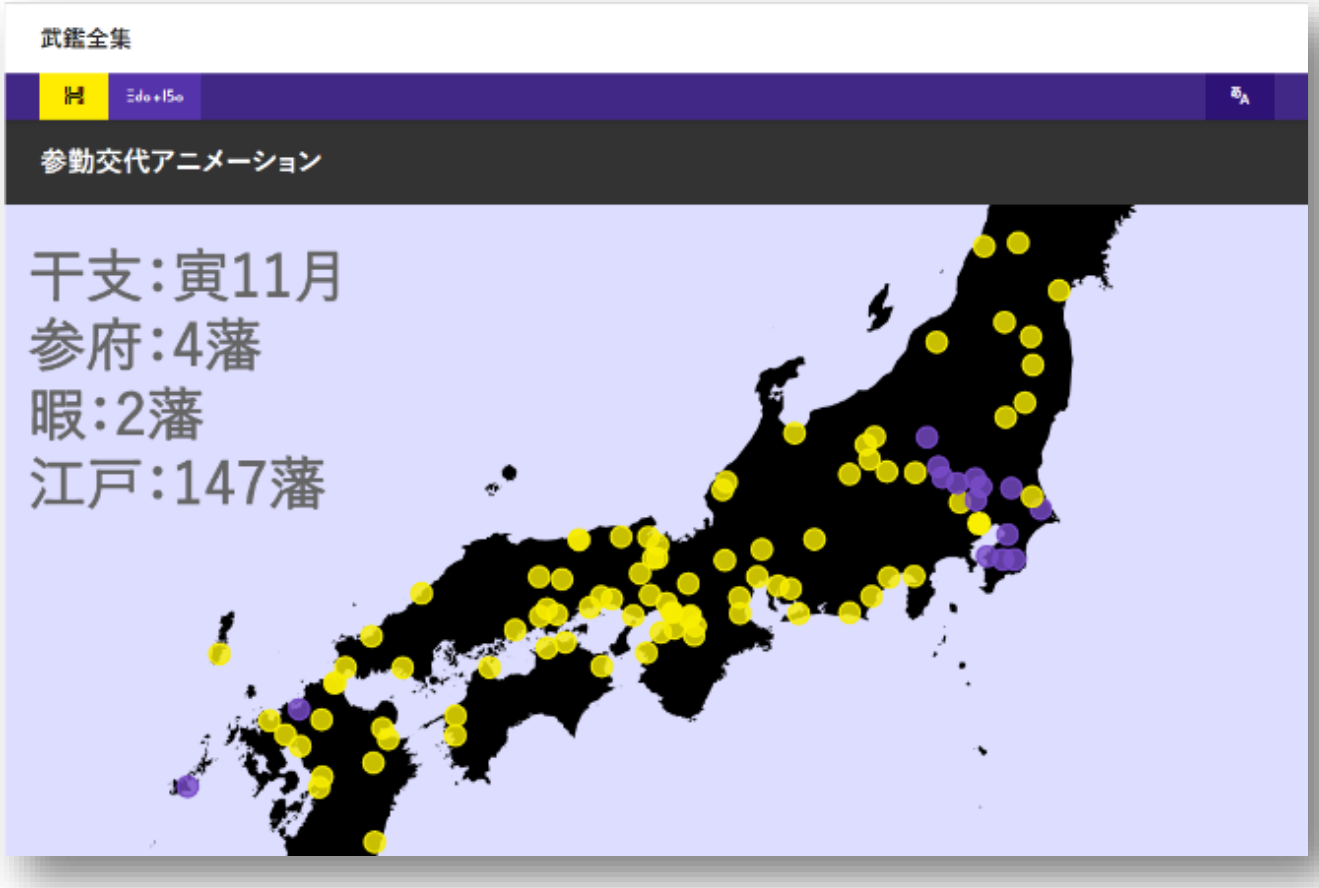
武鑑全集（大名家一覧）

http://codh.rois.ac.jp/bukan/book/200018823/

大名家の基本情報を抽出して一覧表示。

武鑑全集（参勤交代）

http://codh.rois.ac.jp/bukan/book/200018823/trip/



参勤交代時期をアニメーション表示。

江戸情報基盤の構想

- **「日本古典籍データセット」の武鑑：現在8点だが、今月中に373点を追加予定！**
- **「武鑑」の更新情報を用いて、人物、組織、場所などを時間軸に紐付けていく。**
- **「武鑑」を典拠として、他の情報源を紐付けていく（Linked Data）。**
- **任意時期の状況を可視化する時系列インタフェース「江戸タイムマシン」へ。**

新しいリサーチクエスチョン

1. **「武鑑」の網羅的解析**：新しい発見へのブレークスルーになりうる。
2. **出版**：「武鑑」の間違いはどの程度の速さで修正されたのか？
3. **時系列解析**：大名家や幕府役人の人事異動頻度の長期変動パターンは？
4. **画像ベース差分検出**だけで解明？→時系列史料に新たな価値が生まれる。

オープンサイエンスの可能性

- **標準技術としてのIIIF＋オープンなライセンス**：日本に存在する「武鑑」の画像データを、仮想的に集約しながら網羅的な解析を実現。
- **市民科学との連携**：「みんなで翻刻」など、学習と組み合わせた活動の展開。
- **ふるさと翻刻**：故郷の大名に関する情報は、その地方の人々が構造化を担当。

おわりに

まとめ

1. 「武鑑」を時間的に連続して変化する**時系列史料**として定義。
2. 板本書誌学の概念をバージョンと対応させ、差分を読む「**差読**」を提案。
3. **画像ベース差分検出**を用いた**人機分業ワークフロー**による差分翻刻を提案。
4. 「**武鑑全集**」を構築し、将来の網羅的解析に向けた予備調査を実施。

謝辞・参考情報

- **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。**
- **武鑑全集**
  - http://codh.rois.ac.jp/bukan/
- **日本古典籍データセット「武鑑」リスト**
  - http://codh.rois.ac.jp/pmjt/book/?**武鑑**
- IIIF Curation Viewer
  - http://codh.rois.ac.jp/software/iiif-curation-viewer/
- Edo+150
  - http://codh.rois.ac.jp/edo+150/