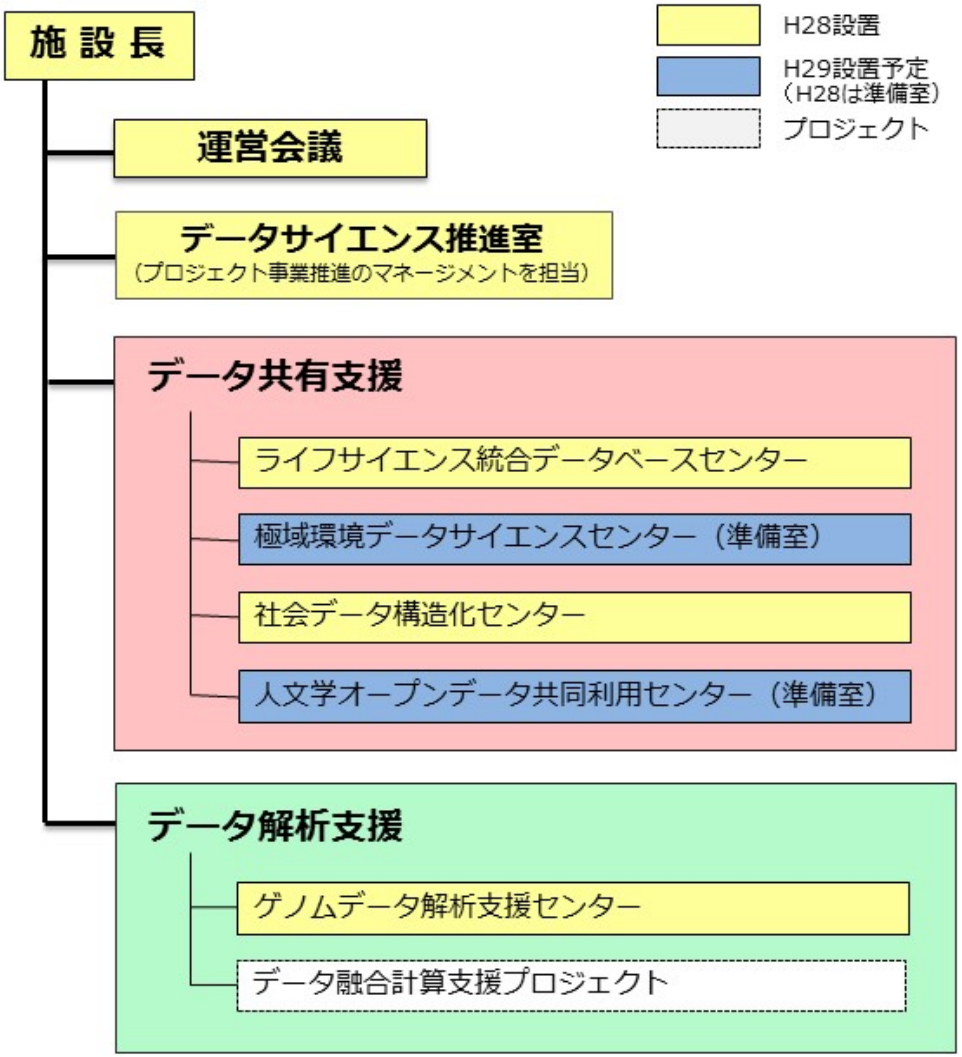
人文学オープンデータ共同利用センターにおける日本語歴史的典籍の利活用

北本 朝展（きたもと あさのぶ）
国立情報学研究所／総合研究大学院大学
人文学オープンデータ共同利用センター
http://agora.ex.nii.ac.jp/~kitamoto/

人文学オープンデータ共同利用センター
Center for Open Data in the Humanities (CODH)

- 情報・システム研究機構のデータサイエンス共同利用基盤施設の一部。
- 2016年4月に準備室開設。2017年4月からセンターとなる予定。
- 国立情報学研究所が主体となり、統計数理研究所と共同して推進。

データサイエンス共同利用基盤施設（組織図）



◆ライフサイエンス統合データベースセンター◆
生命科学分野のオープンサイエンス推進、ライフサイエンス・データベース統合化のための研究開発を推進
（JST/NBDCとの共同研究事業）

◆極域環境データサイエンスセンター◆
過去から現在に至る長大な時間軸を持った極域環境変動・地球システム変動に関する貴重なデータと、その分析・解析支援を提供する共同利用を推進

◆社会データ構造化センター◆
大学研究者のための社会調査データ、公的調査ミクロデータ、ソーシャルビッグデータに関するデータベースを整備。また、データ利用コミュニティを形成し、環境、治安、経済を含む各種の社会的課題の解決のための実証的研究を促進

◆人文学オープンデータ共同利用センター◆
データサイエンスに基づく人文学（人文情報学）という新たな学問分野を創生するとともに、データを中心としたオープン化を推進することで、組織の枠を超えた研究拠点を形成・強化

◆ゲノムデータ解析支援センター◆
最先端のバイオインフォマティクス技術を駆使して大量のゲノム・トランスクリプトームデータから生物学的に重要な情報を得るためのデータ解析支援

◆データ融合計算支援プロジェクト◆
データ融合計算技術による諸科学・産業界での課題解決

https://ds.rois.ac.jp/contents/

オープンデータの価値
1. 利用
2. 透明性
3. 参加

人文系大学研究者

人間文化研究機構

海外研究機関

非人文系研究機関（情報学・統計学）

市民協力者・オープンデータ利用企業

研究者に対する公平なアクセス

異分野研究者の参入

市民・企業等との協働

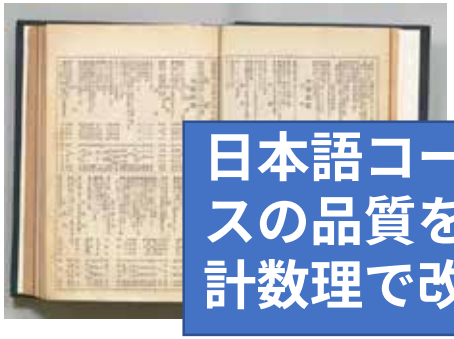
人文学オープンデータ共同利用センター

課題1: データ利用基盤構築

課題2:内容分析

AIはくずし字・古典籍を解読できるか？

課題3:質的向上



日本語コーパスの品質を統計数理で改善

課題4:オープン化

モバイルアプリ等を用いた市民科学

センターの目的・方向性

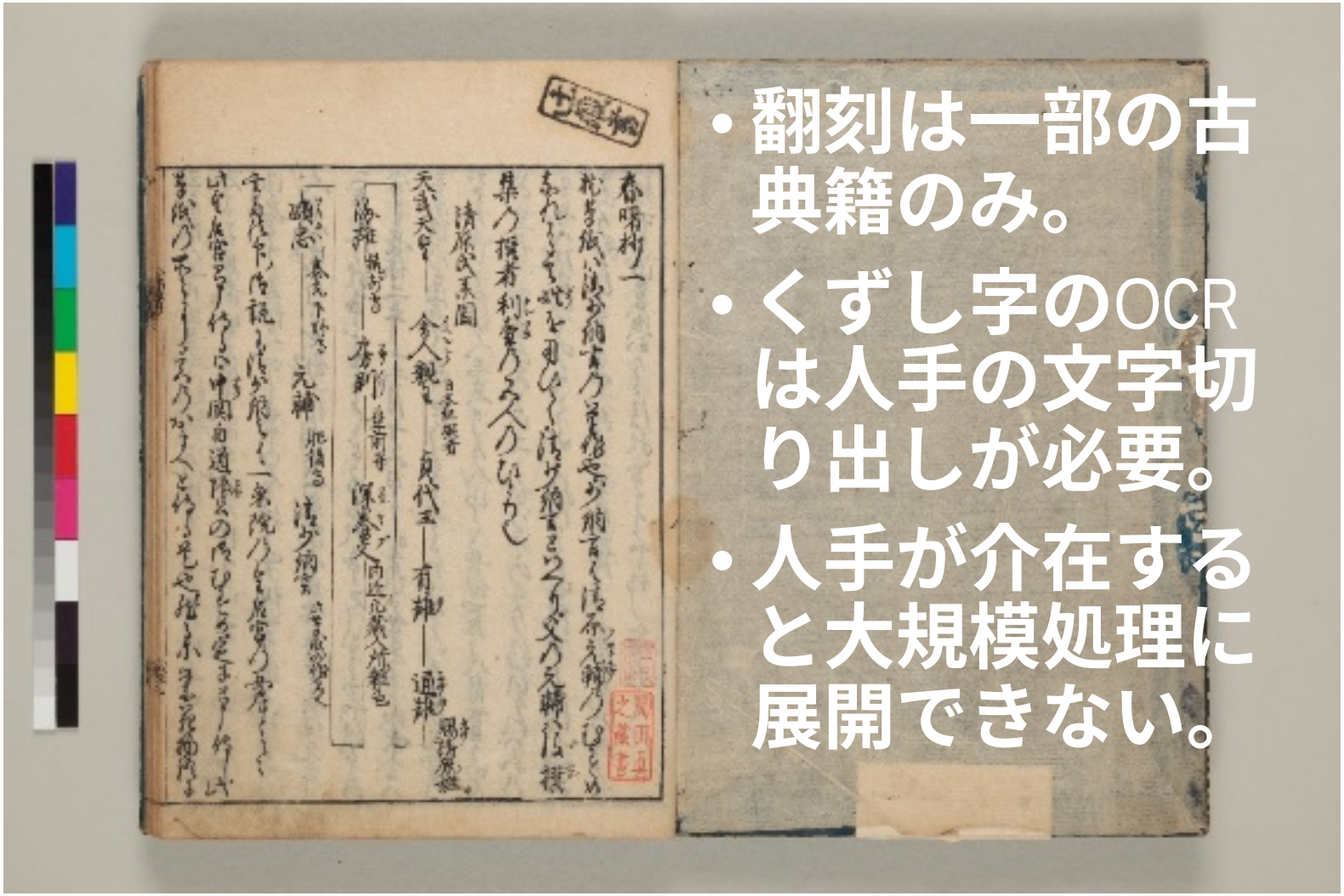
- データサイエンスに基づく人文学（人文情報学）という新たな学問分野を創生するとともに、データを中心としたオープン化を推進することで、組織の枠を超えた研究拠点を形成・強化
- 情報学・統計学の最新技術に基づき、内容分析に基づく「深いデータ公開」を追究。
- 機構間連携や海外機関連携を活用し、日本の人文知を世界に向けて集約、利用、発信。
- オープンデータに基づくシチズンサイエンスやオープンイノベーションの実例を一般化。

歴史的典籍NW事業との連携

人文学オープンデータ共同利用センターの中では国文研は最重要連携先の一つ。

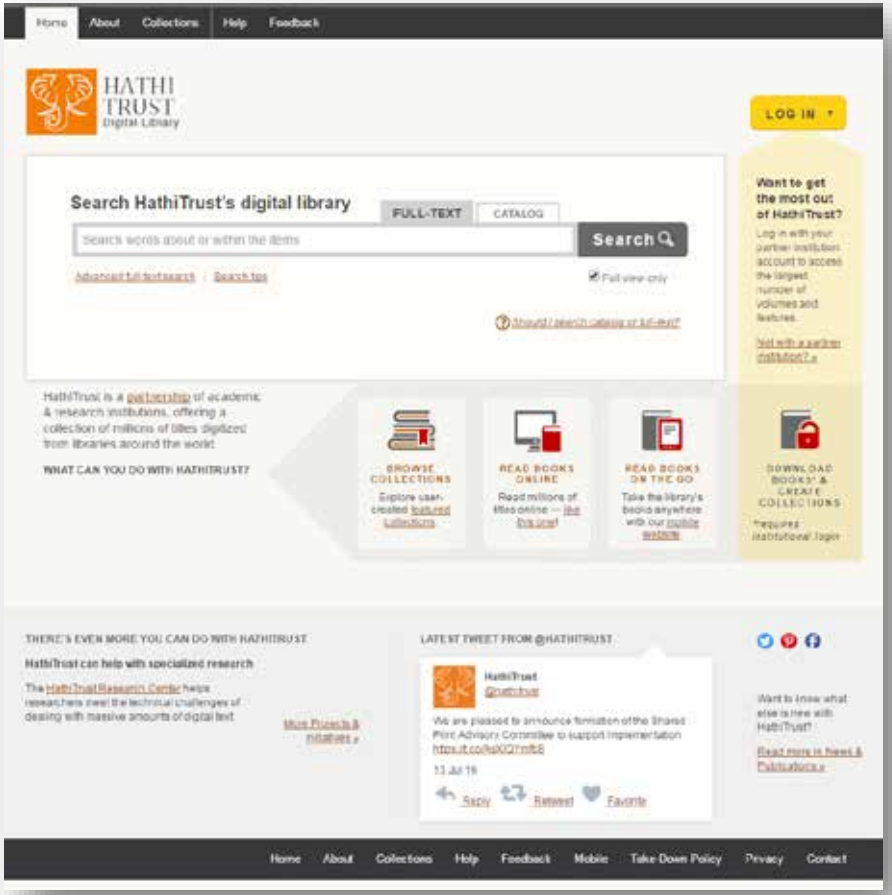
1. **データのオープン化における基礎的条件となる、研究基盤の構築を推進。**
2. **文字情報分析と非文字情報分析の両面から、内容解析に基づくオープン化を開拓。**
3. **関係者の協働に基づくオープン化を、料理本などを例として展開。**

文字情報の網羅的解析



- 翻刻は一部の古典籍のみ。
- くずし字のOCRは人手の文字切り出しが必要。
- 人手が介在すると大規模処理に展開できない。

HathiTrust Digital Library

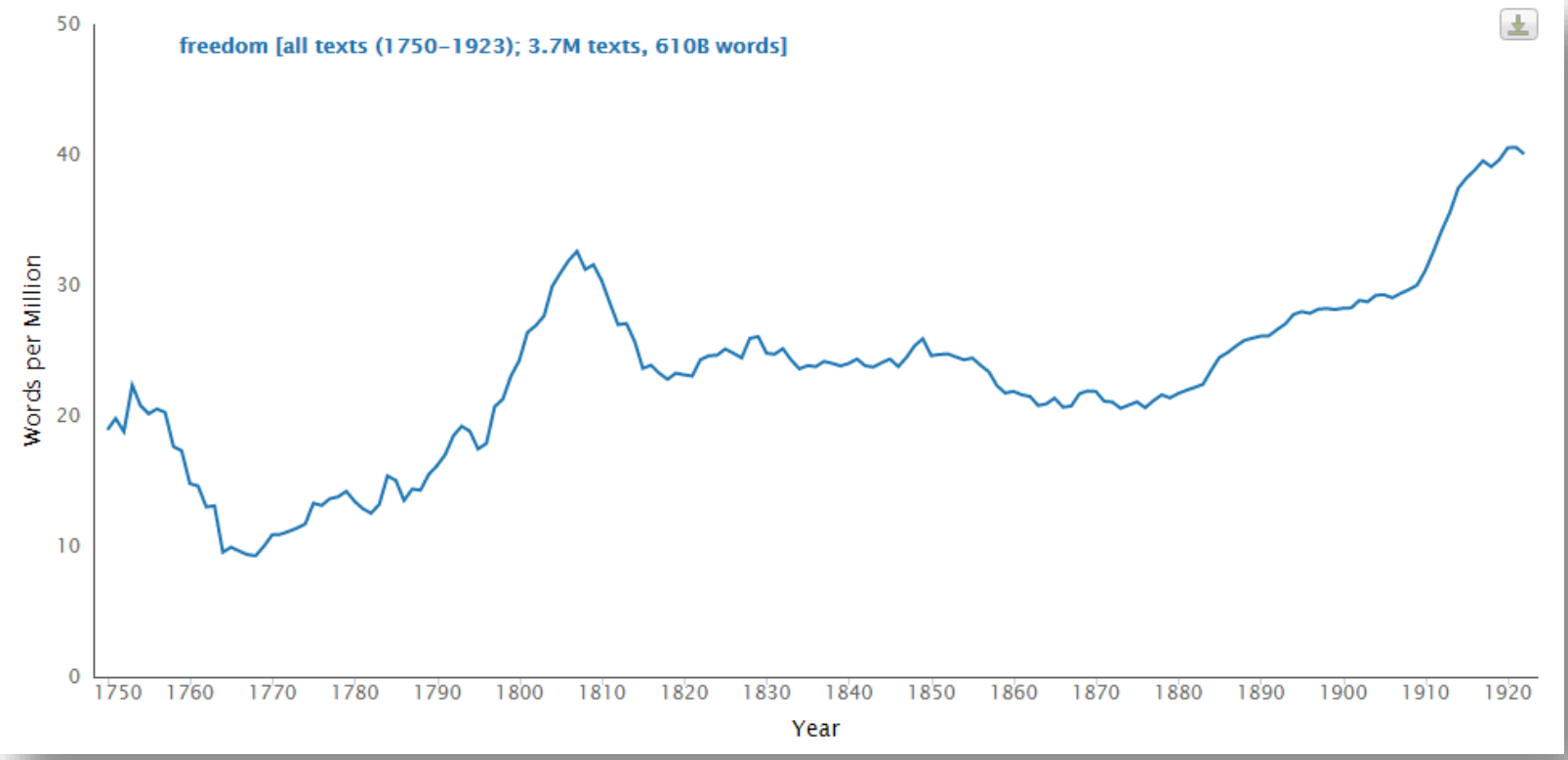


https://www.hathitrust.org/

- **欧米系言語（英語が50%）が中心の書籍1465万冊（51億ページ）をデジタル化。**
- HathiTrust Bookwormでは、460万冊のOCRに基づき、単語の出現頻度グラフを表示（より大規模なGoogle Ngram Viewerもある）。

HathiTrust Bookworm

OCR技術が開拓する、文化の網羅的な解析



https://bookworm.htrc.illinois.edu/

日本文化の網羅的解析

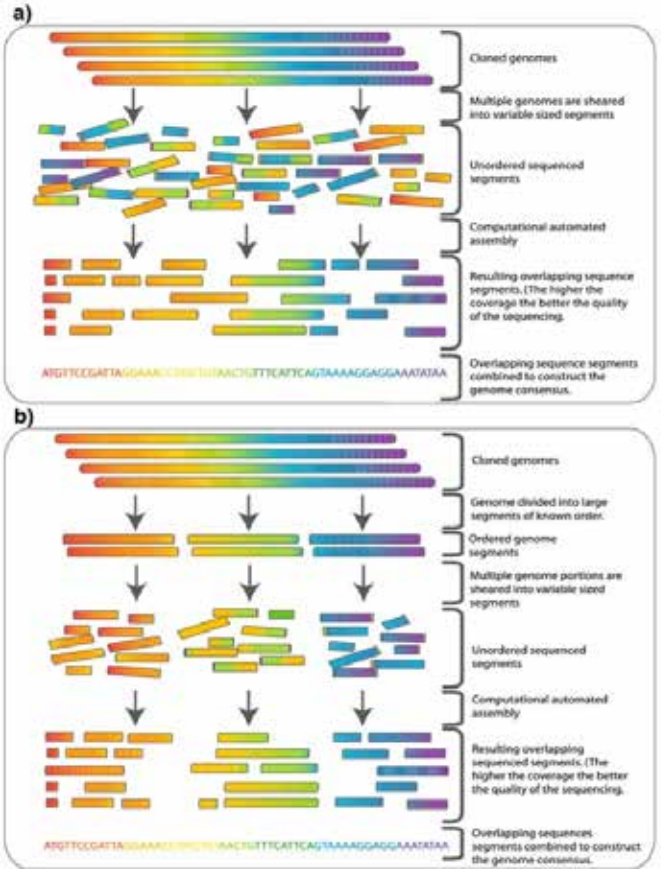
- **近代の書籍**：活字のOCRは長年の蓄積があるが、状態が悪いとまだ難しい。
- **近世の書籍**：くずし字のOCRは人手の介在が必要で、全自動化への道は険しい。
- **技術の急速な発展**：不可能に思えたことが、近未来には可能になることもある。
- **ヒトゲノム解析**：ヒト遺伝子の網羅的解析は、不可能が可能になった例の一つ。

ゲノム解析の歴史

年代	できごと
1953	DNA二重らせんモデルの提唱。
1980年代	ヒトゲノムの全解読には100年かかる？
1987ごろ	日本人研究者が、自動解読による高速化というアイデアを提案。
2003年ごろ	ヒトゲノムの解読が完了。13年間の期間と30億ドルの費用を要した。
2016年	ヒトゲノムの解読は10万円（ただし装置は数千万円）。近い将来には1時間、1000円で可能（装置も数百万円）？

我々の現在地？

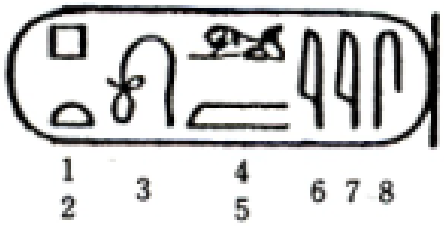
断片を読んでつなげる



- 全体をいったん細かい断片に切り刻む。
- 断片の遺伝子配列（文字列）を解読する。
- **文字列の重なりを活用して断片をつなげる。**
- このアルゴリズム開発が、ゲノム解読の鍵。

ヒエログリフ解読

「プトレマイオス」のカルトゥーシュ



http://gc.sfc.keio.ac.jp/class/2008_26698/slides/04/25.html

- ロゼッタストーン上の王様の名前（カルトゥーシュ）を解読できたことが突破口に。
- **読めるところから読み、エビデンスをつなぎ合わせていく**。
- 長い文字列を断片的に読み、文脈を駆使して徐々に空白を埋めるアプローチはあるか？

スクリプトーム解析

- **一次元の記号列**：遺伝子配列も古代文字も日本語歴史的典籍も、すべて記号列。
- **断片の再構成**：翻刻等の伝統的な読み方とは異なる、新しい読み方は可能か？
- **機械学習**：ディープラーニングは、現代の解読における突破口になりうる。
- **コミュニティ**：機械学習だけでは不十分で、多くのアルゴリズム開発が必要。

くずし字のオープンデータ化

文字データのオープン化の例 MNISTデータベース

https://www2.warwick.ac.uk/fac/cross_fac/complexity/study/msc_and_phd/co902/2013_2014/resources/

- 機械学習の研究を振興するには、学習データの提供が必須である。
- 今年秋には国文研と共同で、字形データセットのオープン化を予定。
- 情報学研究者の参入を促し、新しいアルゴリズムをオープンに開発。

非文字情報の分析

- **書誌学的な情報**：文字情報の内容に立ち入らず、外形的に書籍の特徴を分析する。
- **版・刷の分析**：画像マッチングにより、類似画像の相違を定量的に特定する。
- **絵画的な情報**：絵や図などビジュアルな内容を、アルゴリズム的に分析する。
- **自動タギング**：意味的な単位でタグを付与したいが、線画は抽象的で難しい。

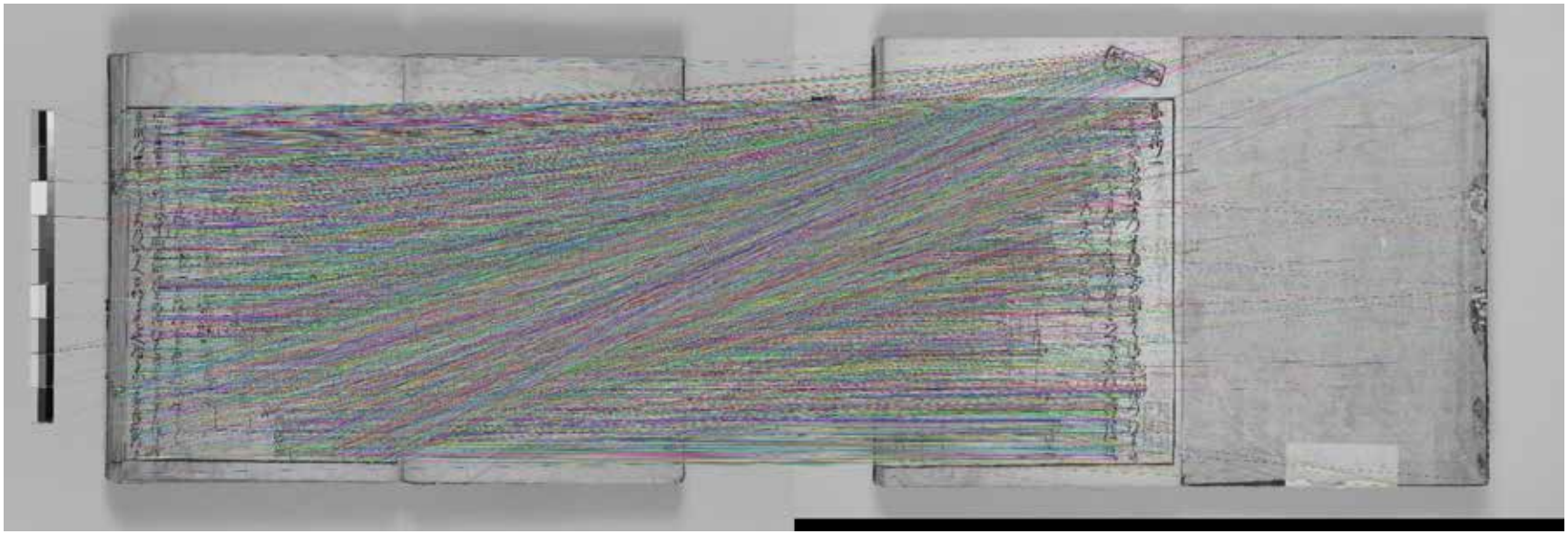
版の分析

春曙抄グループA1
枕草子春曙抄, **国文研高乗**,
089-0338-00002

春曙抄グループA1
春曙抄, **国文研鵜飼**,
096-0815-00002

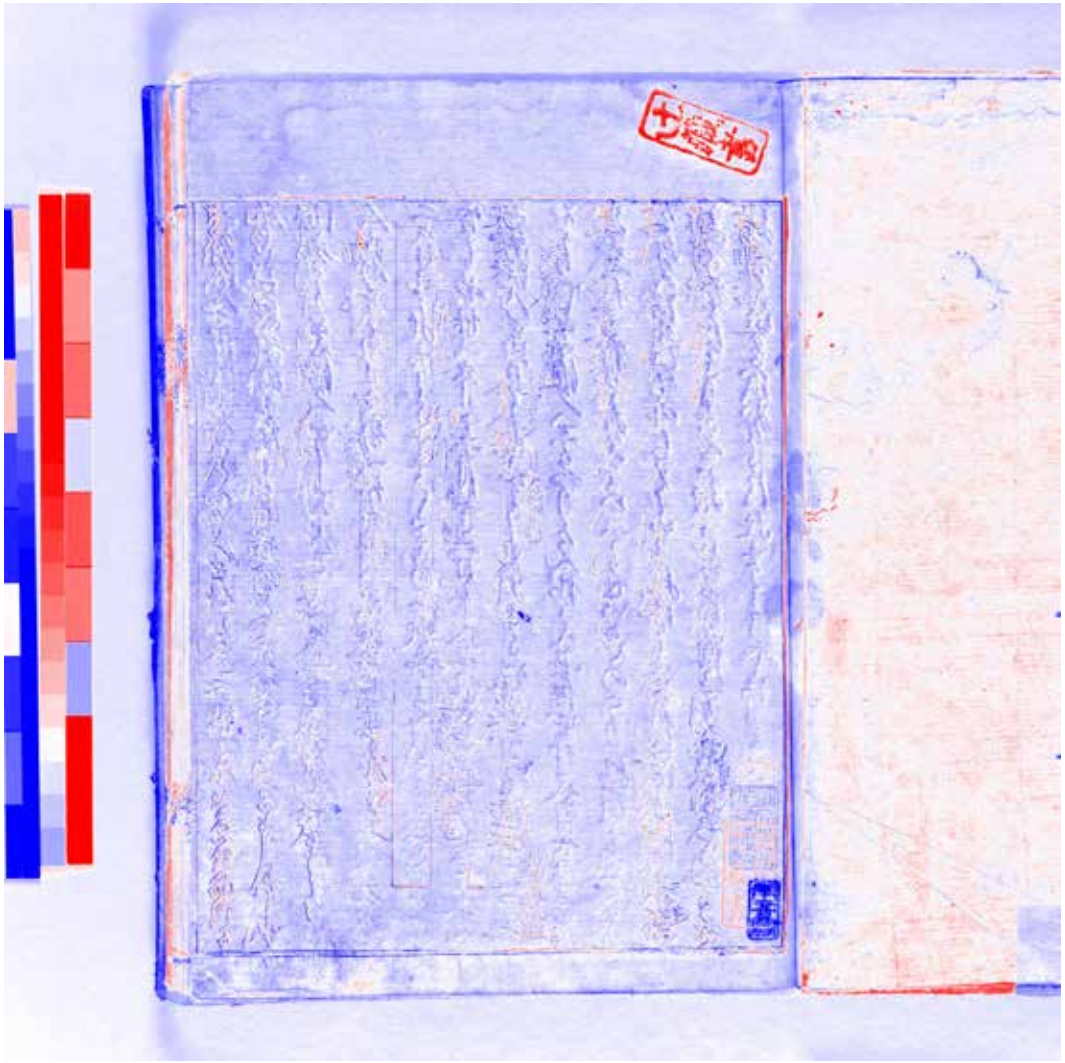
両者の違いを確認する作業は、人間にとっては退屈な仕事…

画像マッチング



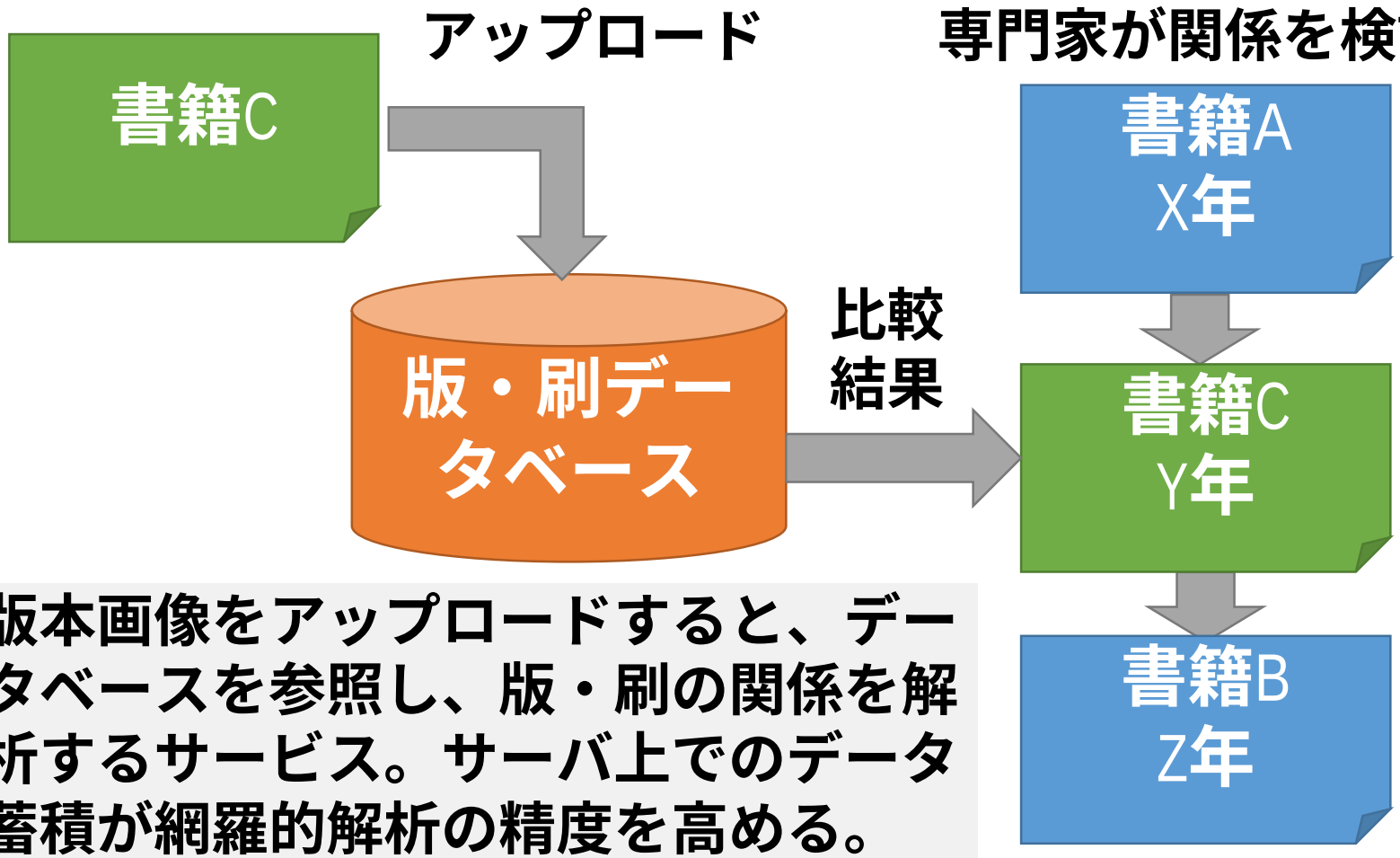
オープンソースソフトウェアOpenCVを用いた画像マッチングにより、両者の対応点を特定し、画像変換パラメータを推定。

カラー強調比較



- 089の方が黒いところを赤、096の方が黒いところを青。
- 文字の部分は大きな違いがなく、両者は同じであると考えられる。
- 印は両者の違いが明確に見える。

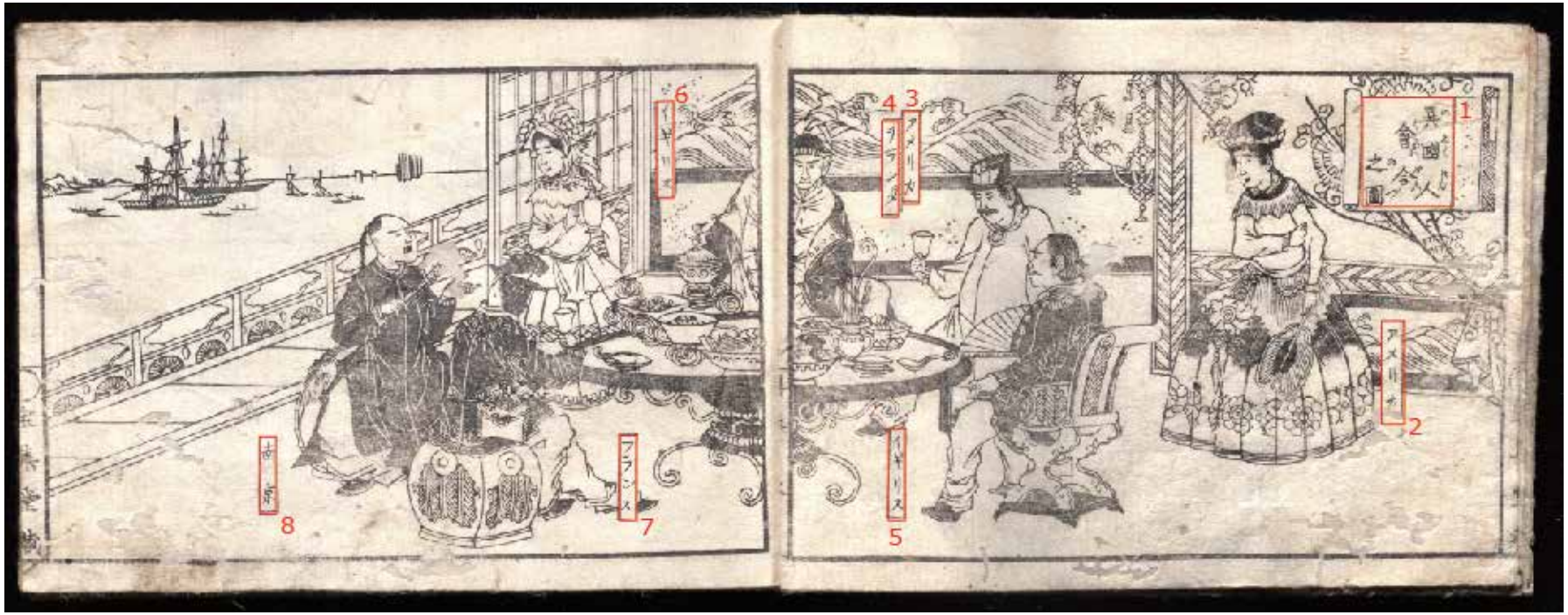
書誌学的解析サービス構想



版本画像をアップロードすると、データベースを参照し、版・刷の関係を解析するサービス。サーバ上でのデータ蓄積が網羅的解析の精度を高める。

生活における利活用

江戸時代の料理はどのようなものか？江戸の料理本のレシピは現代でもおいしいか？



新編異国料理, NIJL0079-049-0228-00005, **国文研古典籍データセット**

市民科学（シチズンサイエンス）

- **研究を市民と共に進めていく方法論。**
- 研究データを市民に提供し、市民が自らの力でデータを増強する。
- 学びの場に市民が参加することで、市民が新たな知識を得る。

http://cookpad.com/kitchen/14604664

まとめ

- **人文学オープンデータ共同利用センター**は、情報学・統計学の概念や手法を活用し、人文学の方法論や可能性を広げることを目指します。
- **歴史的典籍NW事業**では、内容分析や協働を通して、新しい利活用を開拓します。
- **人材募集、共同研究募集**などを開始する予定です。皆様からのコンタクトをお待ちしております！

関連情報

- **データサイエンス共同利用基盤施設**
  - https://ds.rois.ac.jp/
- **人文学オープンデータ共同利用センター（仮サイト）**
  - http://agora.ex.nii.ac.jp/codh/
- **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 第6号**, pp. 4-5, 2016**年6月**
  - http://www.nijl.ac.jp/pages/cijproject/images/fumi_6.pdf