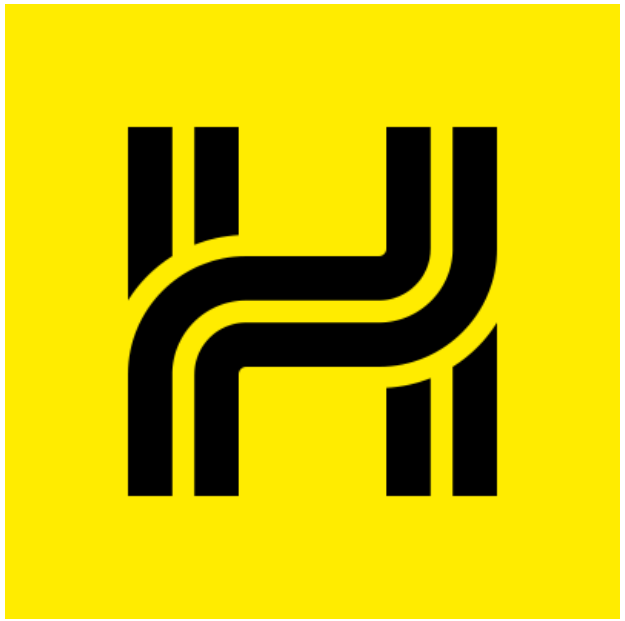
デジタル人文学への招待ーAI、共同研究、デジタル変革



北本 朝展（Asanobu KITAMOTO）

ROIS-DS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

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

https://researchmap.jp/kitamoto/

自己紹介

https://researchmap.jp/kitamoto/

@kitamotoasanobu

- 北本朝展（きたもと あさのぶ）
- 国立情報学研究所 教授、ROIS-DS人文学オープンデータ共同利用センター センター長
- 専門は、情報学、デジタル・ヒューマニティーズ、データ駆動型サイエンス（地球科学・防災等）など。オープンサイエンスの展開に向けた、超学際的な共同研究の展開にも取り組む。

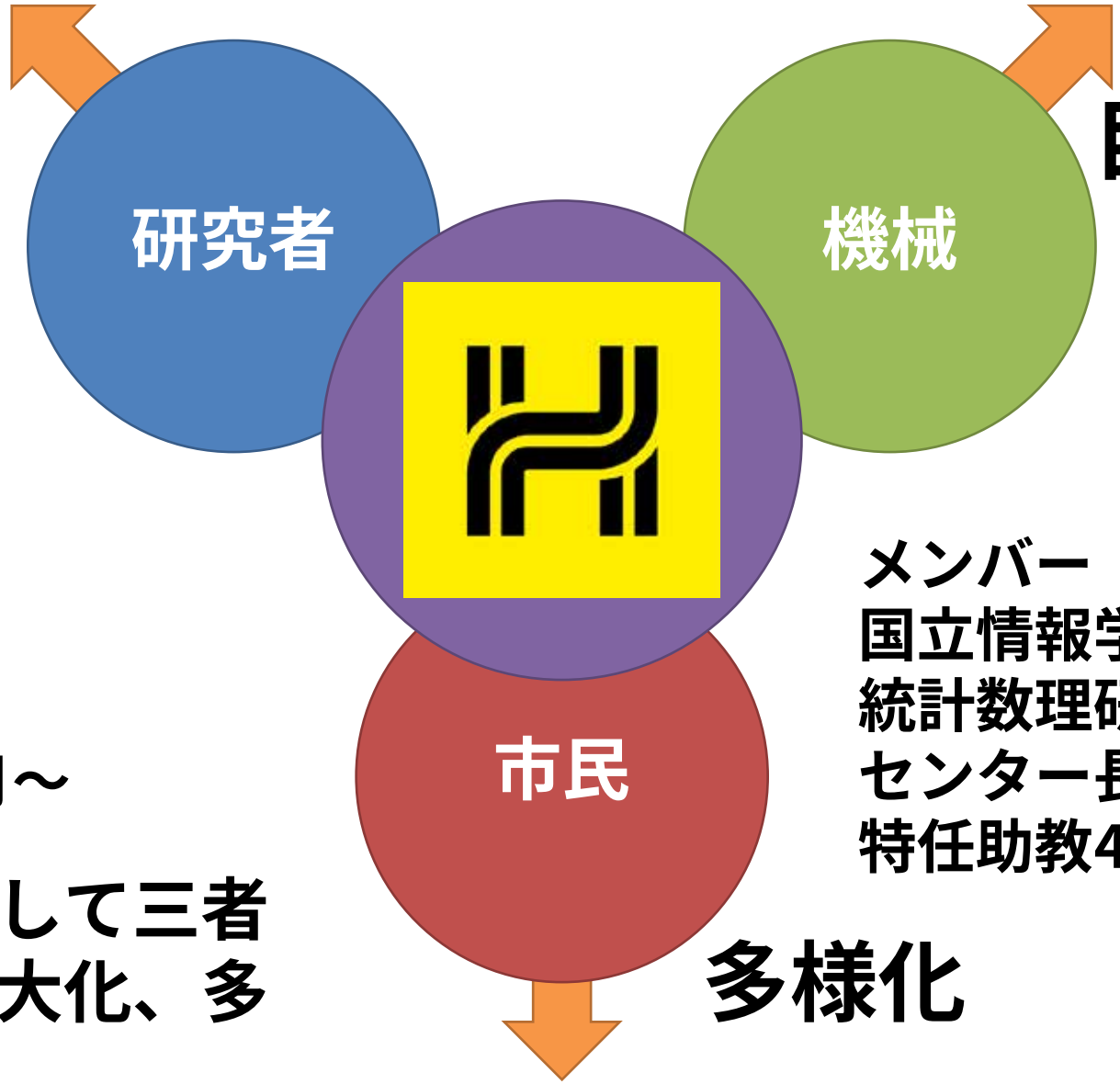
ROIS-DS人文学オープンデータ共同利用センター（CODH）

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

2016年4月～

深化

巨大化



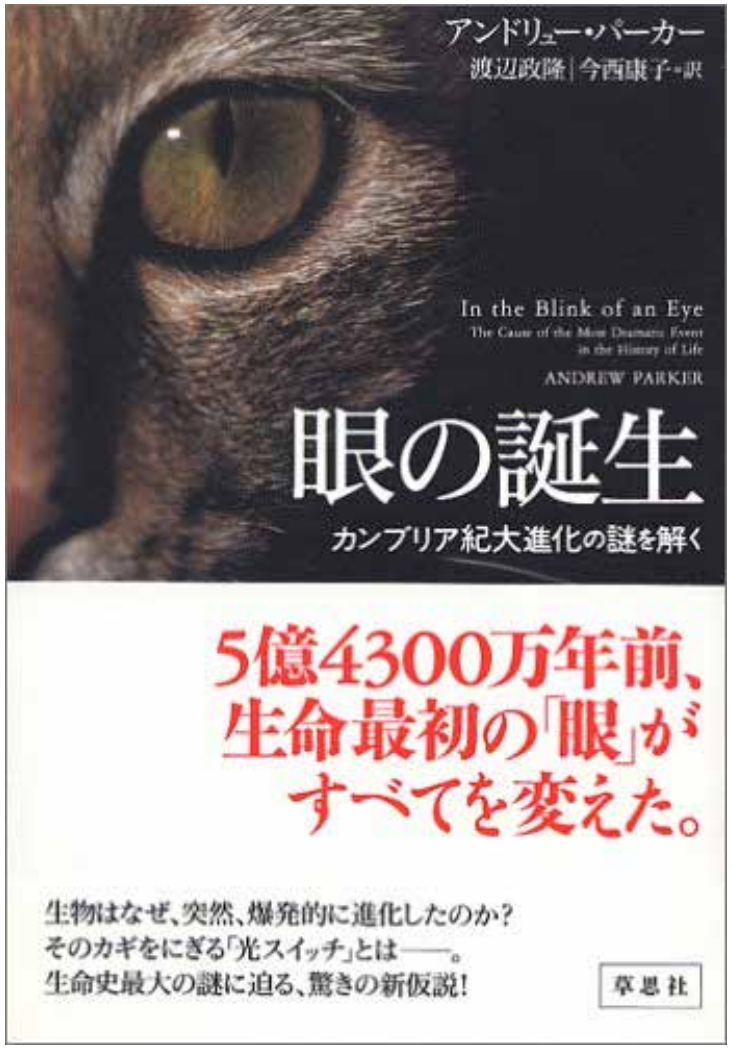
メンバー
国立情報学研究所
統計数理研究所
センター長＋
特任助教4名

多様化

「オープン」の概念を核として三者を接続し、知識の深化、巨大化、多様化を目指す

デジタル人文学とは？

1. **デジタル技術を人文学研究に導入することで、人文学の新しい研究方法を考案し、新しい知識を得ること。**
2. **デジタル技術の一つが「人工知能（AI）」。人間と機械の分業（協力）により、研究方法の変革が起こる。**
3. **コロナ禍が明らかにしたこと＝リモートでアクセスし、議論し、共有できることの価値も大きい。**
4. **共同研究（チーム）が重要性を増す＝「ワンオペ」では研究資源が足りないため、協働の文化が鍵を握る。**



眼の誕生――カンブリア紀大進化の謎を解く, アンドリュー・パーカー,草思社, 2006

AIによる新しい研究方法

- AIという新しいツールによって、**「見えるもの」が変わる**。
- **遠読**（distant reading）：大量のテキストを機械が読み込み、データからパターンを見出す。
- Google BooksとN-gram viewer：数百年にわたる語の使用頻度の推移をグラフ化し、社会の変化を探る。

オープンサイエンス

IIIF Curation Platform
http://codh.rois.ac.jp/icp/

1. **オープンデータ**を公開することで、許可なく、誰でも、データを活用した研究ができる。
2. **オープンソースソフトウェア**によって、無料で、改造可能な研究の基盤が使えるようになる。
3. **超学際的な研究者の協働、市民参加（支援）**などによって、従来型共同研究の枠を超えて、技術や知識の多様性を高める。

どのように共同研究するか？

人文学者と情報学者では、関心のありかが異なる！異文化への理解と、ゴールを共有できる相手が必要。

1. **人文学者のリサーチクエスチョン**をきっかけとして、情報学者がデータ化、システム化の手法を練っていく。
2. **情報学者の技術的提案**をきっかけに、人文学者が自分の研究への活用を進め、システムの課題を出していく。
3. **人文学者と情報学者がアイデアを議論**しながら、新しい研究課題と技術的な解決策を探していく。

篆書字体データセットの事例

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

人文学者

情報系に近い人文学者

人文系に近い情報学者

情報学者

人文学者（国文研 青田 寿美先生）のプロジェクトに、情報学者の立場から、データセット構築に協力。

教訓：分野間の橋渡し人材がいないと、協働はスムーズに進まない。

くずし字データセットの利活用例

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

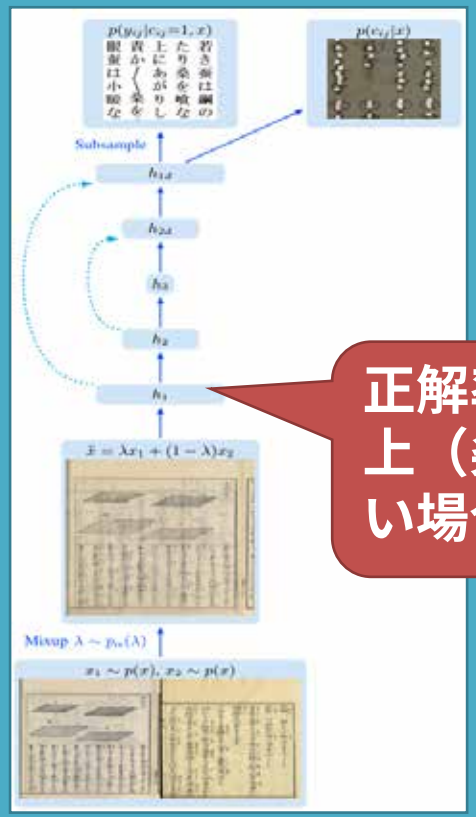
日本古典籍データセット（国文研蔵）

くずし字データセット（国文研・CODH作成）

file	char	x	y
200003803_00024_2.jpg	U+3067	416	114
200003804_00024_2.jpg	U+3055	232	115
200003805_00024_2.jpg	U+304A	327	115
200003806_00024_2.jpg	U+3068	145	116
200003807_00024_2.jpg	U+3046	369	116
200003808_00024_2.jpg	U+305F	457	116
200003809_00024_2.jpg	U+5FA1	104	117
200003810_00024_2.jpg	U+3072	191	118
200003811_00024_2.jpg	U+540D	279	120
200003812_00024_2.jpg	U+3061	501	120

CODH カラーヌワット・タリンほか

KuroNet
AIくずし字認識



正解率95%以上（条件がよい場合）

くずし字認識サービス

kaggle

くずし字認識コンペ

AIくずし字認識アプリ「みを」

くずし字データセットの事例

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

1. **国文研のNIJL-NWプロジェクトでは、古典籍のデジタル化に加え、テキストの全文検索も構想していた。**
2. **くずし字データセットを構築し、AIくずし字認識（機械学習）を開発する計画も始まっていた。**
3. **当初想定のデータ形式では機械学習の可能性が狭まることに気づき、情報学者として仕様変更を主張した。**
4. **この仕様変更は、その後のKuroNetの開発やKaggleコンペの開催において、決定的に重要な役割を果たした。**

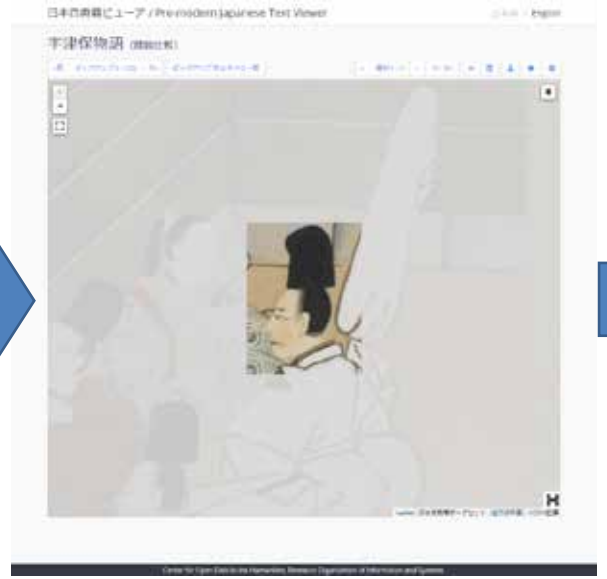
教訓：情報学者がデータの仕様策定段階から入るべき。

顔コレデータセットの利活用例

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

日本古典籍データセット（国文研蔵ほか）

IIIF Curation Viewer / IIIF Curation Platform（CODH開発）



顔貌コレクション・データセット（CODH作成）

美術史研究

AI自動顔認識

機械の創造性（GAN）

CODH **鈴木 親彦、東大 髙岸 輝、**Google Yingtao Tian、EPFL Alexis Mermet**ほか**

顔コレデータセットの事例

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

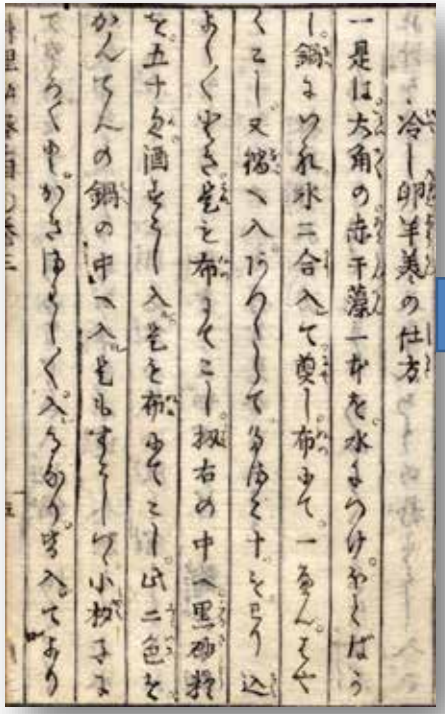
1. CODHでは、IIIF画像を切り取り集める機能を備えた、**オープンソースのIIIF Curation Platformを開発**している。
2. **人文学者（美術史）**が、顔を切り取って集めたら面白いのではないかと考えた。
3. 自らによる作業に加え、**大学院生への謝金も活用し**、数千枚の画像コレクションを構築した。
4. 顔データをオープン化することで、**機械学習研究者が顔認識モデルを開発**し、半自動切り抜きに発展した。

教訓：人文学者が構築した高品質なデータセットは、機械学習研究者が新たな研究を始めるきっかけになる。

江戸料理レシピの利活用例

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

日本古典籍
データセット
（国文研蔵）



データクリエーター、
料理研究家との協働
（CODH主導）

翻刻

現代語訳

料理研究家
によるレシ
ピ化＋料理
写真撮影

江戸料理レシピ・データ
セット（CODH作成）

料理レシピサイト

デパートイベント

鶏鳴新聞
keimei.ne.jp

業界新聞特集記事

江戸料理レシピの事例

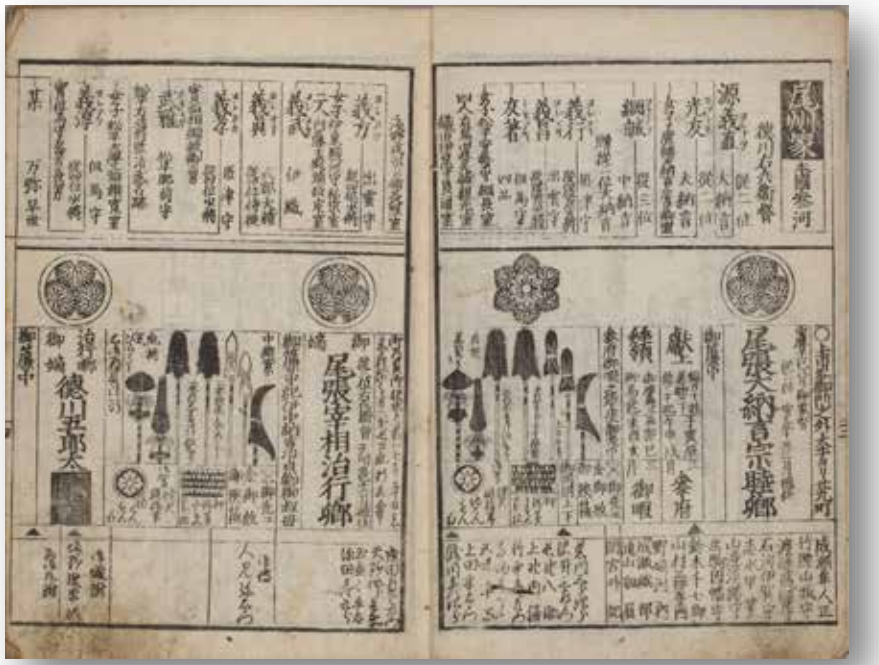
http://codh.rois.ac.jp/edo-cooking/

1. **国文研が主催するアイデアソン**に、情報学者である北本が参加して、江戸の料理本に触れる。
2. レシピに写真を加え、料理レシピサイトに投稿すれば、**現代の生活と接続できる**のではないかと思いついた。
3. **クックパッド**に連絡して共同研究体制を構築。翻刻とレシピ化が可能な**データクリエーター**に作業依頼。
4. **プロの料理研究家**が参画し、調理可能なレシピ化と写真撮影を進め、現代にも通用する水準のレシピを完成。

教訓：アイデアの現実化には、**多様な専門性を備えたチームの協働が必要である。**

「武鑑」の網羅的な解析

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



寛政武鑑 (1789)、日本古典籍データセット（国文研蔵、CODH公開）
http://codh.rois.ac.jp/pmjt/book/200018823/

1. **江戸の大名や幕臣に関する「データブック」。紋などのパターンも含み、豊富な構造化データを記録。**
2. **実用：役職移動情報、江戸観光みやげ、大名行列観覧ガイド等。**
3. **江戸時代の約200年にわたって出版され続けたベストセラー。**
4. **貴重な大規模データを、デジタル人文学研究に使える形式にしたい。**

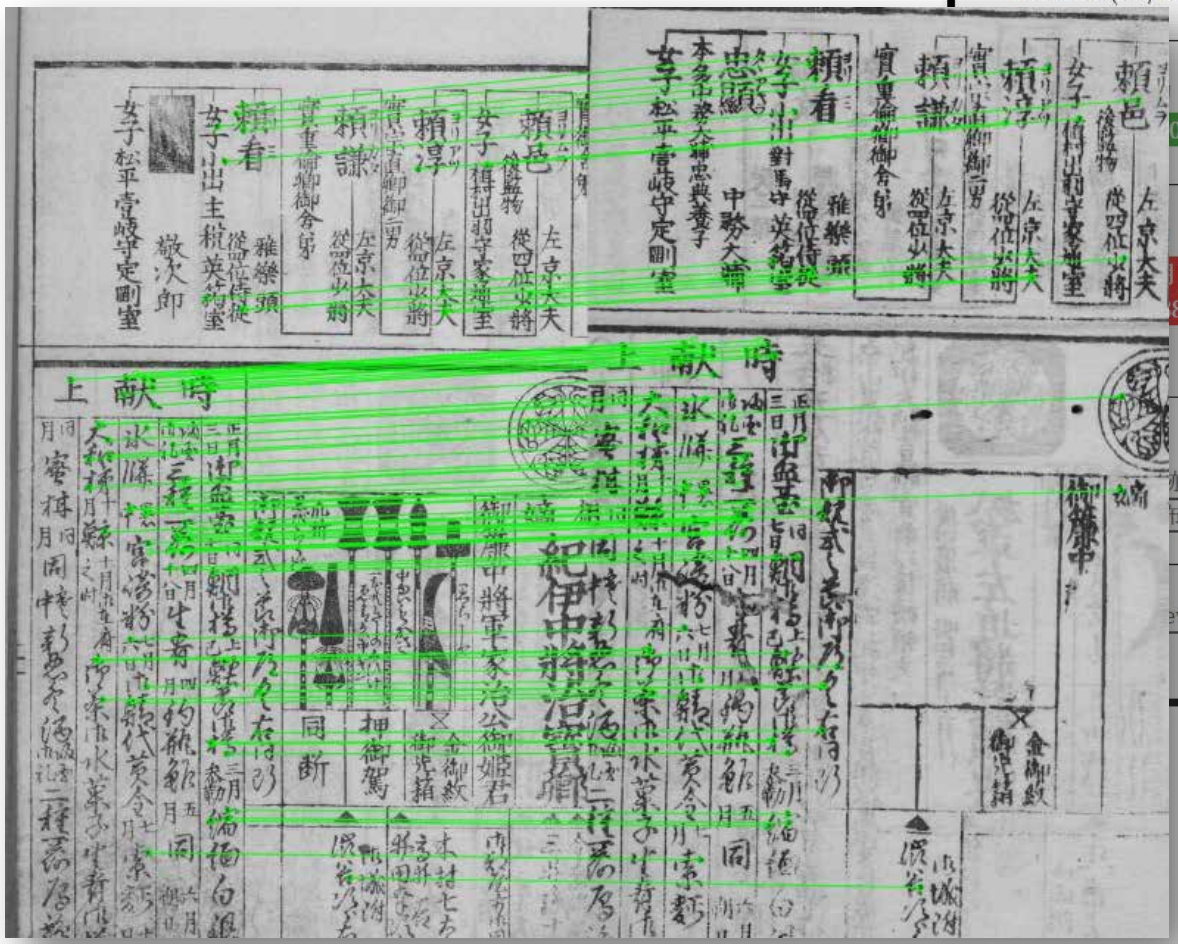
参考：藤實 久美子、江戸の武家名鑑 武鑑と出版競争、吉川弘文館、 2008

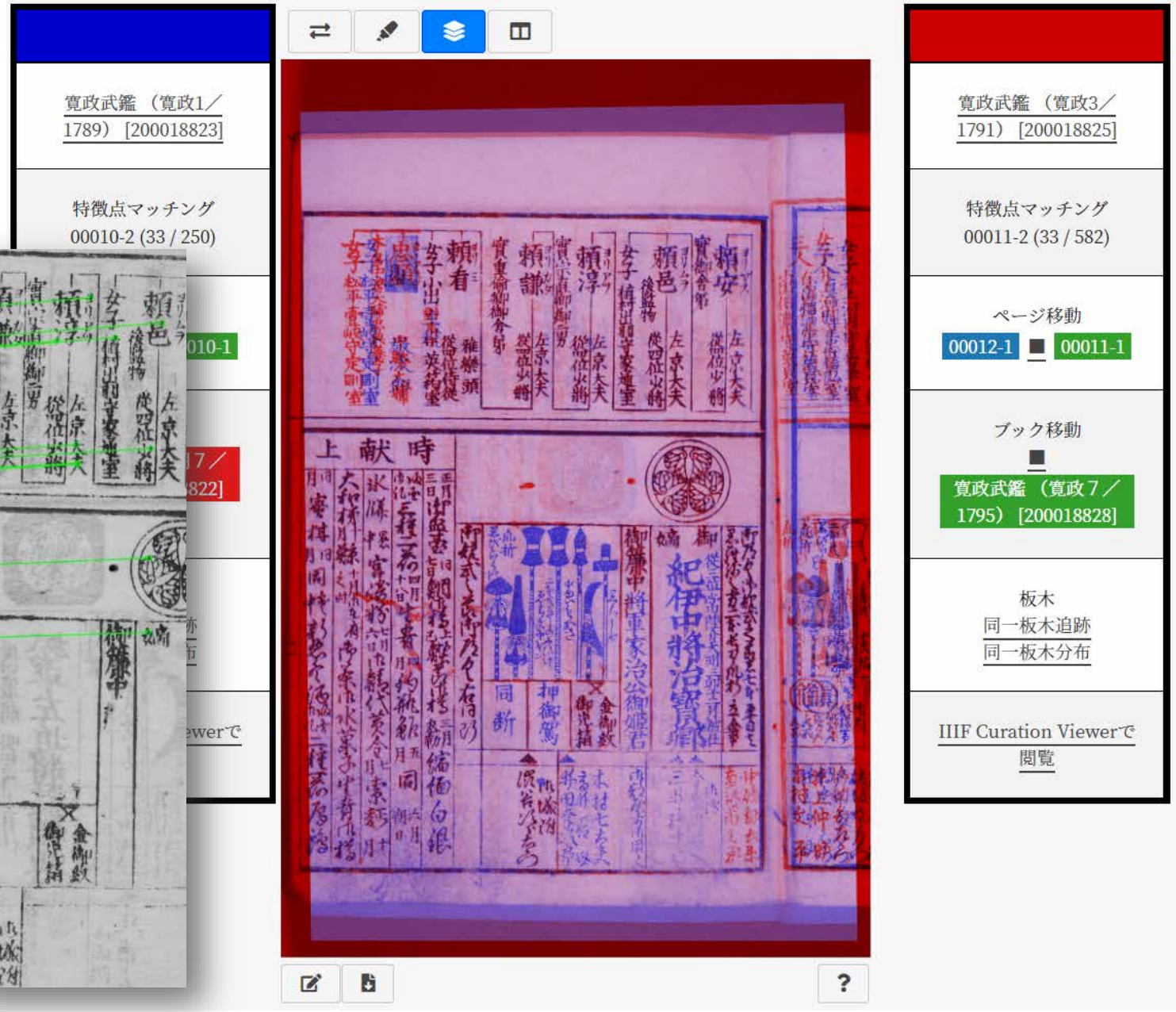
差分翻刻のアイデア

問い：200年にわたる史料をどう翻刻するか？

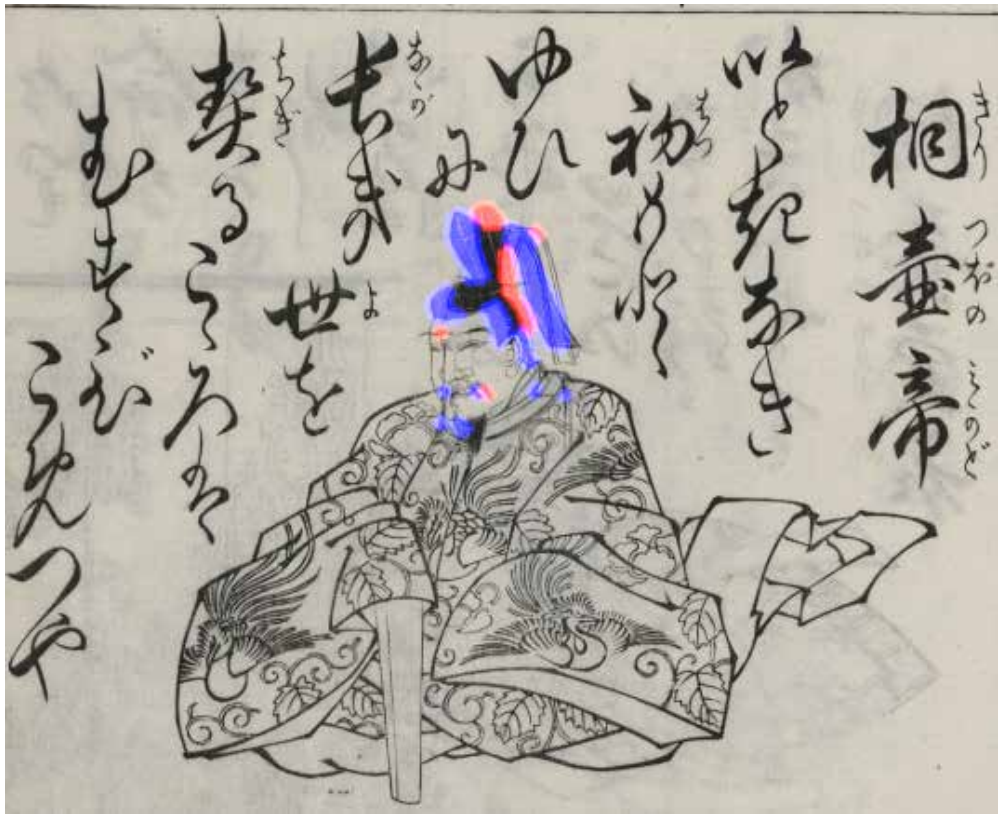
解決：前後の版本で変化した部分のみを翻刻し時系列データを構築する。

ページ照合




寛政武鑑 （寛政1／1789） [200018823]
特徴点マッチング
00010-2 (33 / 250)
寛政武鑑 （寛政3／1791） [200018825]
特徴点マッチング
00011-2 (33 / 582)
ページ移動
00012-1
00011-1
ブック移動
寛政武鑑 （寛政７／1795） [200018828]
板木
同一板木追跡
同一板木分布
IIIF Curation Viewerで閲覧

差読（Differential Reading）のための画像照合サービス http://codh.rois.ac.jp/differential-reading/



源氏百人一首（パタパタ顔比較）、東京大学総合図書館

1. vdiff.jsを活用し、任意の画像を指定し、照合結果を表示・共有できるサービスを公開。
2. ウェブ版＝URLを指定
3. ファイル版＝フォルダを指定
4. 外部サービスと連携可能。

武鑑全集の事例

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

1. 情報学者である北本が武鑑を見たとき、**異なる版の比較問題にコンピュータビジョン技術が使える**ことに気づく。
2. **技術は15年以上前から存在した**が、それが版本の版間差分の強調に使えるという発想が乏しかった。
3. 武鑑研究の第一人者である**藤實久美子教授に相談し共同研究開始**。当時は岡山で勤務していたが、その後国文研に異動。
4. 任意の版本を比較できるプラットフォームへと発展できれば、**書誌学的な研究において画期的な効率向上**が期待できる。

教訓：技術がきっかけの例。問題そのものは昔からあったが、技術の動向を知らないと、適切な解決策は思いつけない。

歴史ビッグデータの統合解析

http://codh.rois.ac.jp/historical-big-data/

過去のビッグデータを統合解析するための基盤技術を研究

歴史的資料（史料）

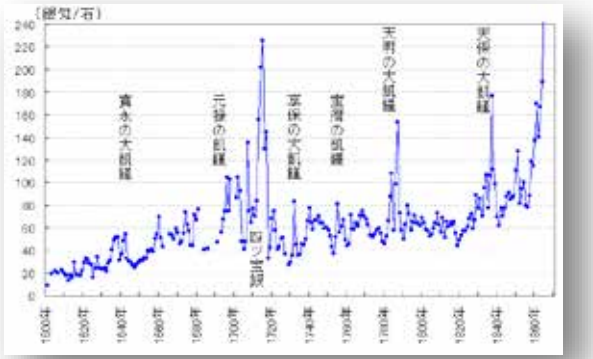
自然科学的データ

人文社会的データ

- 気候
- 地震
- 噴火
- 疫病
- 経済
- 人口
- 政治
- 文化

データ構造化ワークフロー

歴史ビッグデータ研究基盤（機械可読データ）



江戸ビッグデータの統合

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

現代地図との重ね合わせ

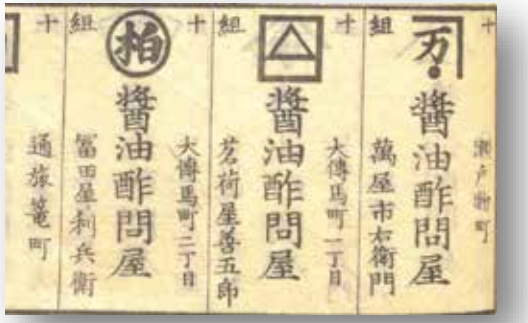
画像：江戸切絵図（国立国会図書館）
地名：CODHが、データクリエイターと共に、IIIF画像への注釈としてデータ整備

CODH **鈴木 親彦ほか**

地名識別子の整備・共有（CODH）

GeoLOD

https://geolod.ex.nii.ac.jp/



商業ビッグデータ

観光ビッグデータ

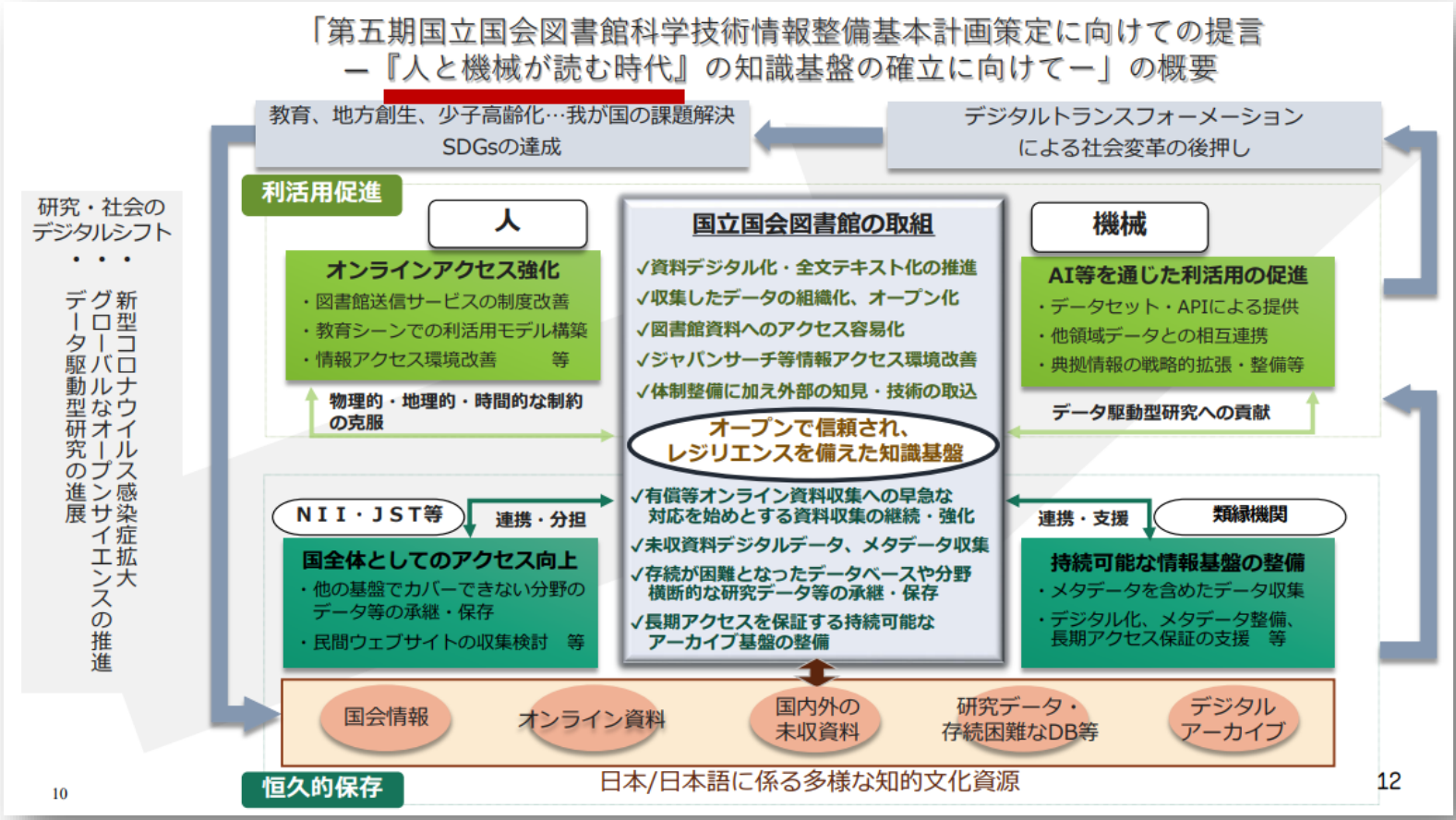
江戸ビッグデータの事例

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

1. **国立国会図書館**などが公開するIIIF画像に対して、CODHがアノテーション（注釈）を加え、**付加価値**を生みだした。
2. IIIFの相互運用性を活用することで、**機関横断的にアーカイブを統合**し、基盤データに新たな**付加価値**を与えられる。
3. GeoLODという地名識別子を活用することで、**異なる分野のデータセットを統合**し、「**総合知**」につなげていく。
4. 地理情報や業種分類、観光地などを**現代と接続**することで、日本文化の資産として活用していく道が開ける。

教訓：相互運用性・識別子などを、機関や分野をまたいで共通化することで、データセットの付加価値はさらに高まる。

図書館の役割



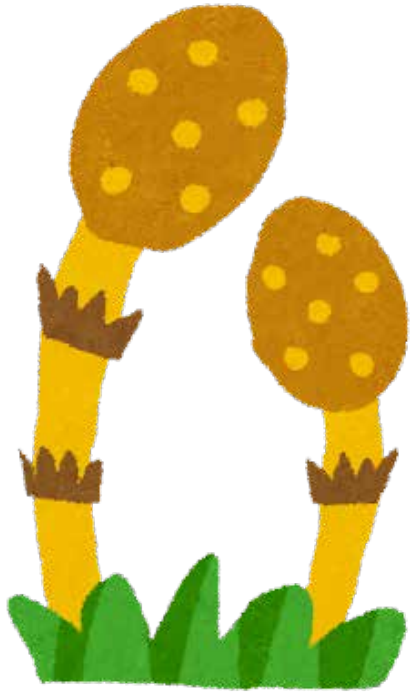
「人と機械が読む時代」の知識基盤の確立に向けて
https://dl.ndl.go.jp/info:ndljp/pid/11631622

1. **図書館の読者として「機械」を想定**しているか？**機械が訪問しやすい図**書館とは何か？
2. **機械の助けによって、図書館の価値**を上げられるか？**全文検索など、新しいサービスを作**れるか？

機械が訪問しやすい図書館とは？

1. **共通の方式でアクセスできる。**
   - **IIIF（International Image Interoperability Framework）など、どこの組織のデータにも共通の方式でアクセスできる。**
2. **使いやすいライセンスが設定されている。**
   - **クリエイティブ・コモンズ（Creative Commons）など、オープンな利用に対応可能な利用許諾方法を提供する。**
3. **識別子などの仕組みで資料に安定してアクセスできる。**
   - **DOI（Digital Object Identifier）など、長期間にわたって、同一の方法で、資料を識別しアクセスできる保証がある。**

機械で価値を増す図書館とは？



可愛いつくしのイラスト
いらすとや

1. **機械を用いて、「人が読む」ための新しいサービスを実現する。**
2. **人間の技術や知識によって、機械を賢くできれば、好循環が回り出す。**

例：日本古典籍（古文書？）に、AI**くずし字認識を適用し、全文検索エンジン「つくし」を実現すれば、古**典籍の研究方法は大きく変わる！

人文学研究とデジタル変革（DX）

AIなどのデジタル技術の活用、従来の枠を超えた共同研究などにより、人文学の研究手法を変革し、これまでにない新しい知識を得る。

1. **機械の役割は「下読み」や「検索」など人間（専門家）の支援。「知識」や「発見」は人間の役割であり、研究者が不要になることはあり得ない。**
2. **機械の支援で新しいものが見えてきた研究者は、そうでない研究者に比べて優位になる可能性がある。**