歴史ビッグデータ：過去の記録の構造化と発見のための新たな方法論



北本 朝展（Kitamoto Asanobu）

ROIS-DS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

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

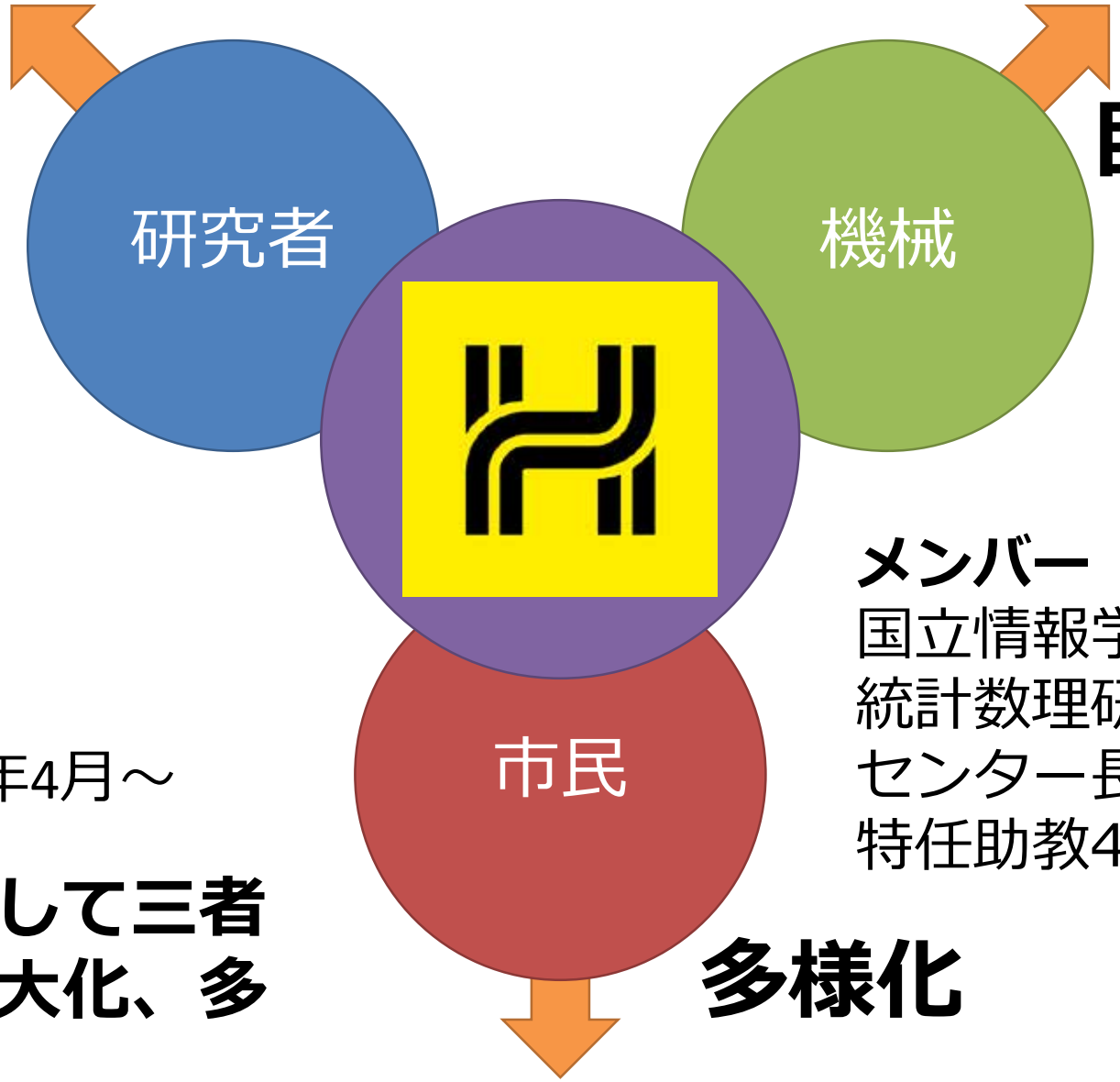
ROIS-DS人文学オープンデータ共同利用センター（CODH）



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

歴史ビッグデータとは？

1. **現代のデータを想定して開発されたビッグデータの方法論を、過去の記録に延長すること。**
2. 生データの構造化：機械可読なデータを生成し、様々な分野における解析などに利用する。
3. 多様な（Variety）データの統合解析：単体のデータだけでは見えてこない新しい知見を得る。
4. 人文学研究のデジタルトランスフォメーション：研究のデジタル化には、技術要素の深化だけでなく、研究体制や（評価）制度の変革も関わる。

歴史ビッグデータの統合解析

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）



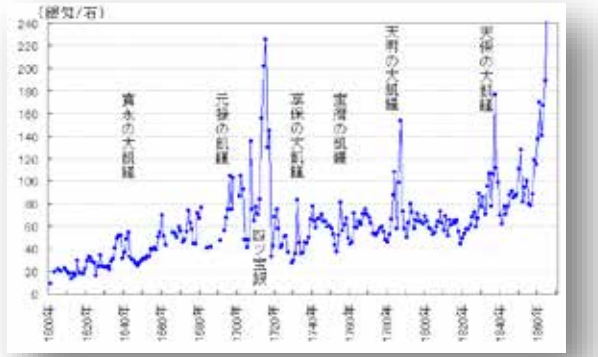
自然科学的データ

人文社会的データ

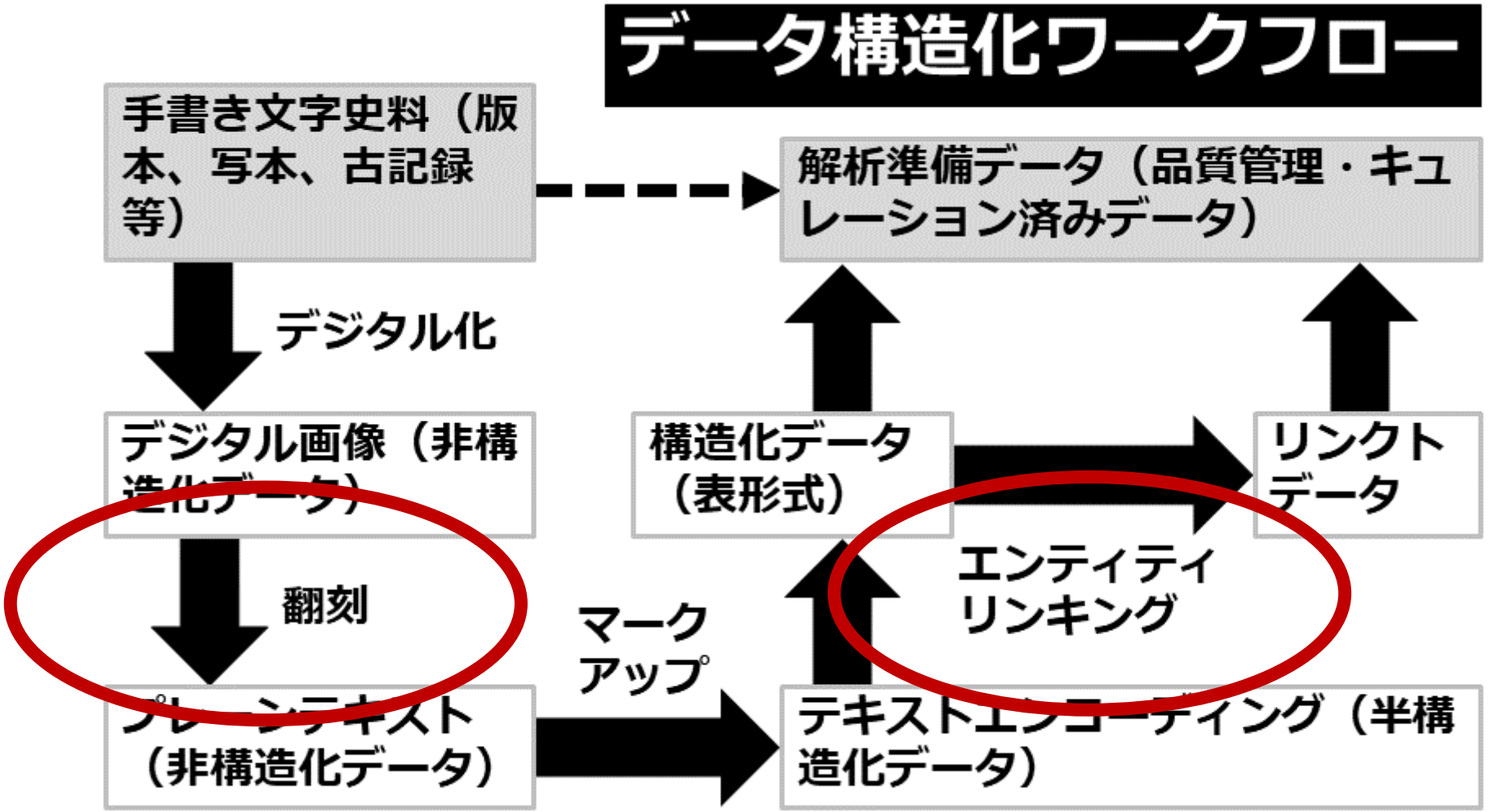
- 気候
- 地震
- 噴火
- 疫病
- 経済
- 人口
- 政治
- 文化

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）



データ構造化ワークフロー



ビッグデータを巡る言葉の整理

データ駆動型人文学

- データを基盤とする新しい手法により、人文学の研究手法を革新する。
- データの共有やエビデンスに基づく検証などの新しい文化を人文学に導入する。
- 人文学的に新しい知識を得ることを目標とする。

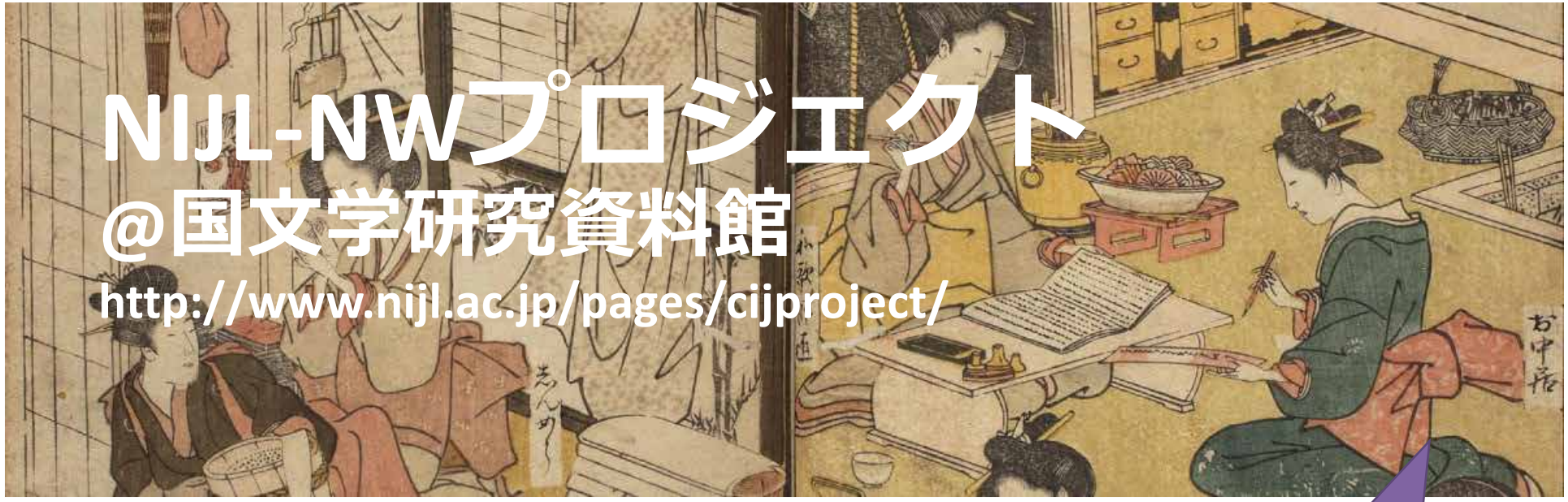
人文学ビッグデータ

- 人文学から生み出される大規模データにより、（非）人文学の研究手法を革新する。
- 人文学データに特有の構造化や解釈などの技術を分野を超えて共有する。
- 定量的またはデータに基づく知識を得ることを目標とする。

今日のセミナーの構成

1. 歴史ビッグデータ：過去の記録の構造化に基づくデータ駆動型人文学研究（北本）→ **歴史ビッグデータのコンセプト**
2. シルクロード遺跡の再発見：東洋文庫アーカイブと遺跡データベースの構築（西村） → **シルクロード歴史研究の事例**
3. みんなで翻刻と歴史地震研究（加納） → **古地震研究の事例**

- 2は人文学研究がデータ駆動型のアプローチでどう変わるかを紹介する。
- 3は人文学分野のビッグデータが非人文学分野にどのようなインパクトを与えるかを紹介する。



NIJL-NWプロジェクト
@国文学研究資料館
http://www.nijl.ac.jp/pages/cijproject/

300,000点の日本古典籍（1868年以前）を、デジタル化し、オープンデータとして公開。

日本文化のビッグデータをどのように活用するか？

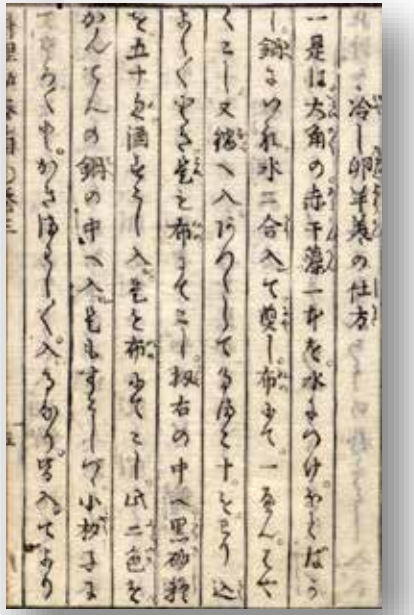
過去の文化遺産と「くずし字」の問題



文書
数億点

日本に残された古典籍や古文書の点数（冊数）の推定

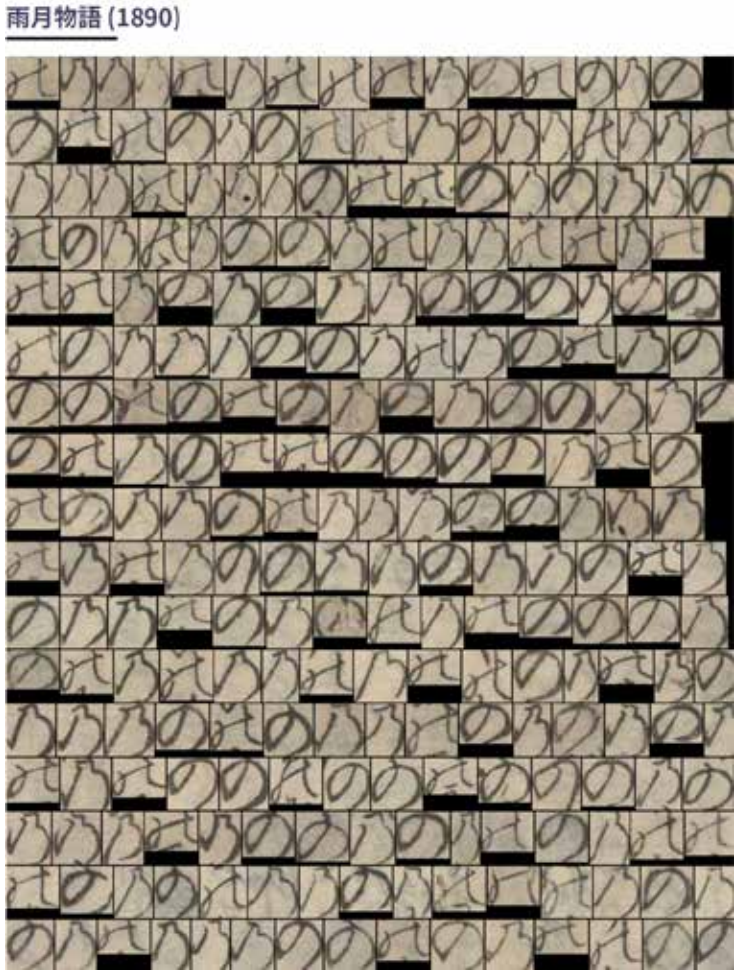
読者
数千人



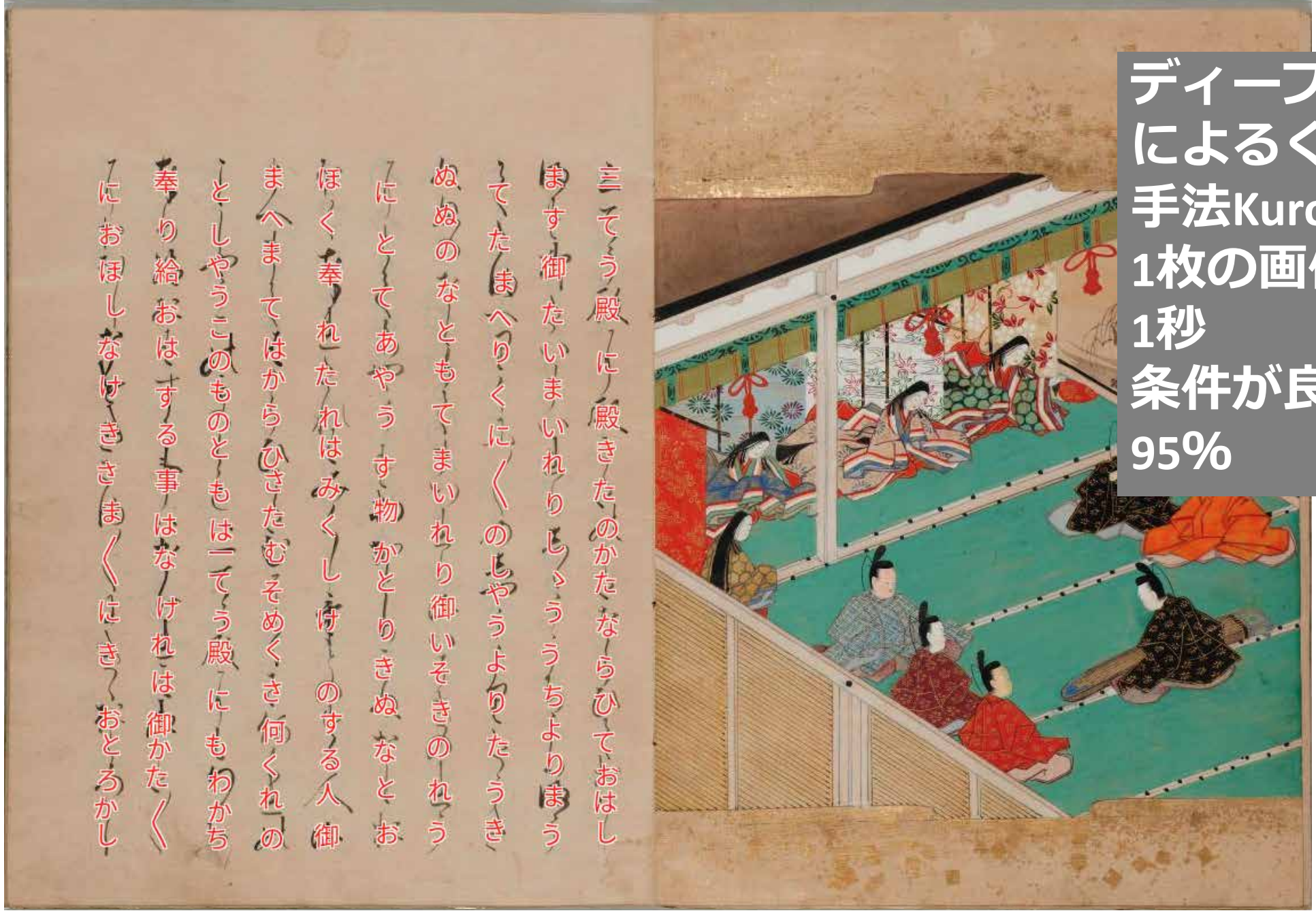
くずし字をきちんと読める人数の推定
（**全人口の0.01%**）

くずし字データセット

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



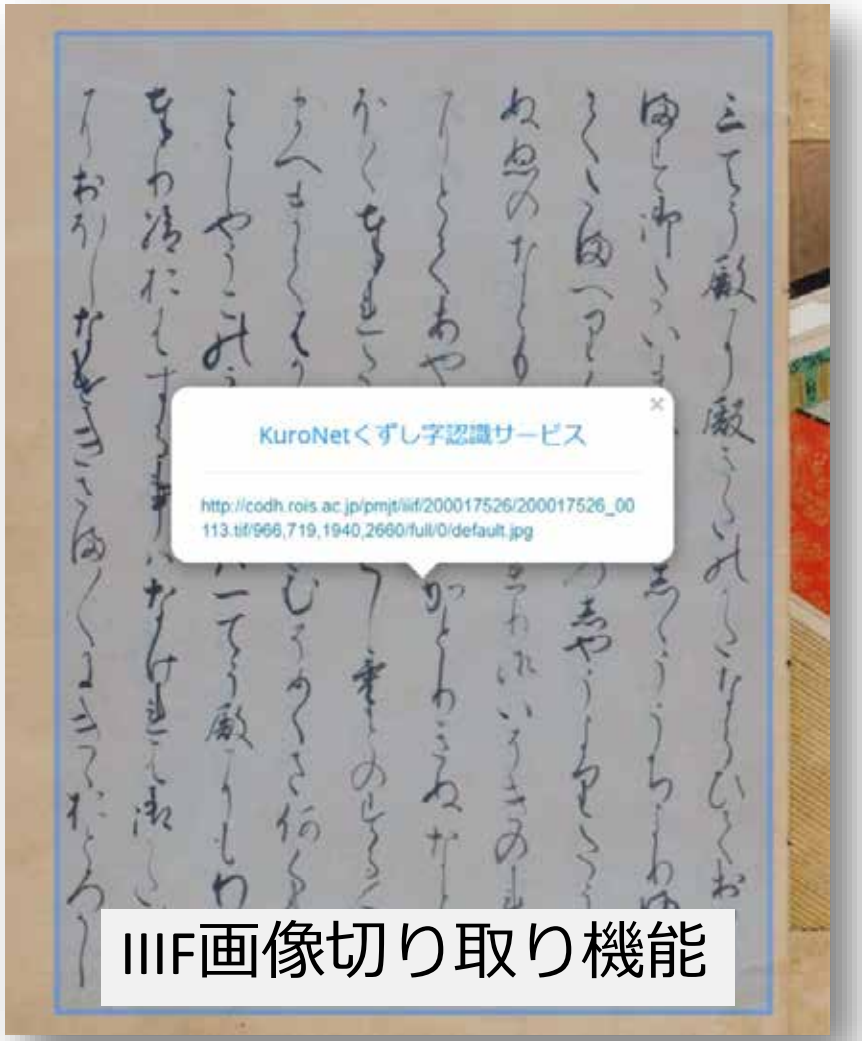
- 国文学研究資料館が作成、CODHが整理して公開するオープンデータ。
- 2020年7月現在
  - 文字種＝4,328
  - 文字数＝1,086,326
- ZIPファイルダウンロード→くずし字認識の訓練データとして利用。
- データセットの公開が、AIによるくずし字認識研究を活性化した。



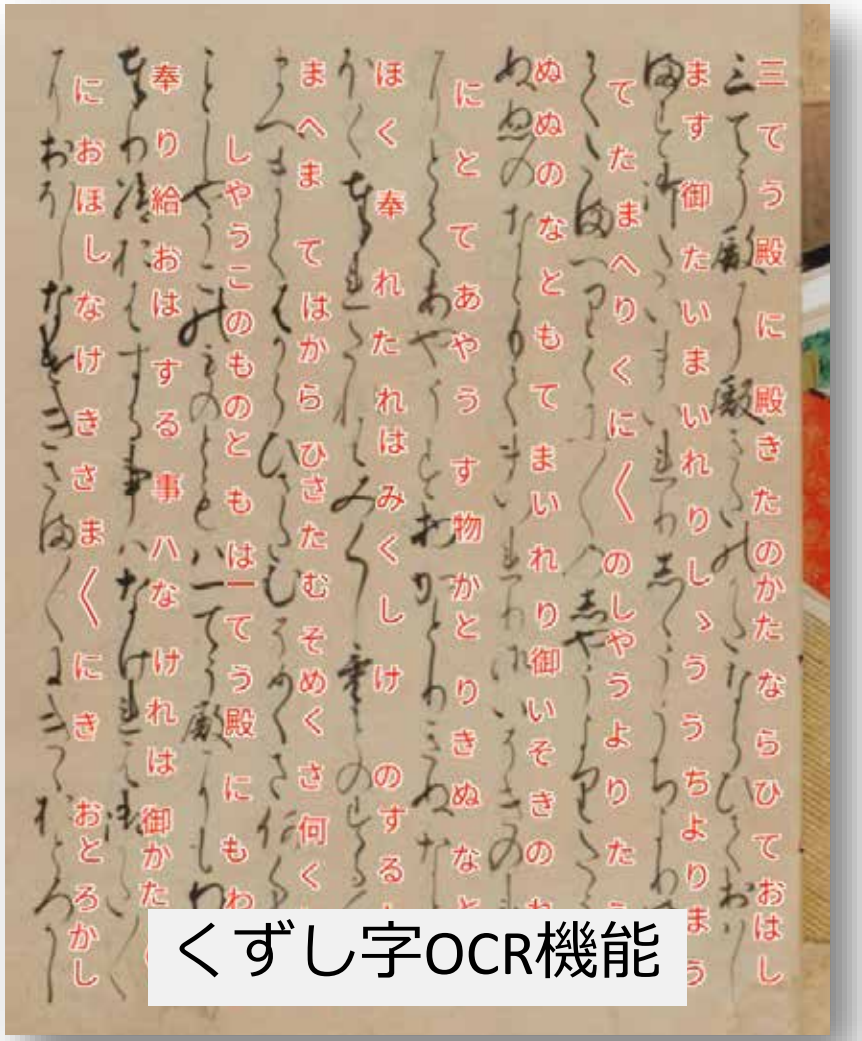
ディープラーニングによるくずし字認識手法KuroNetを開発
1枚の画像の認識に約1秒
条件が良ければ精度95%

KuroNetくずし字認識サービス

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



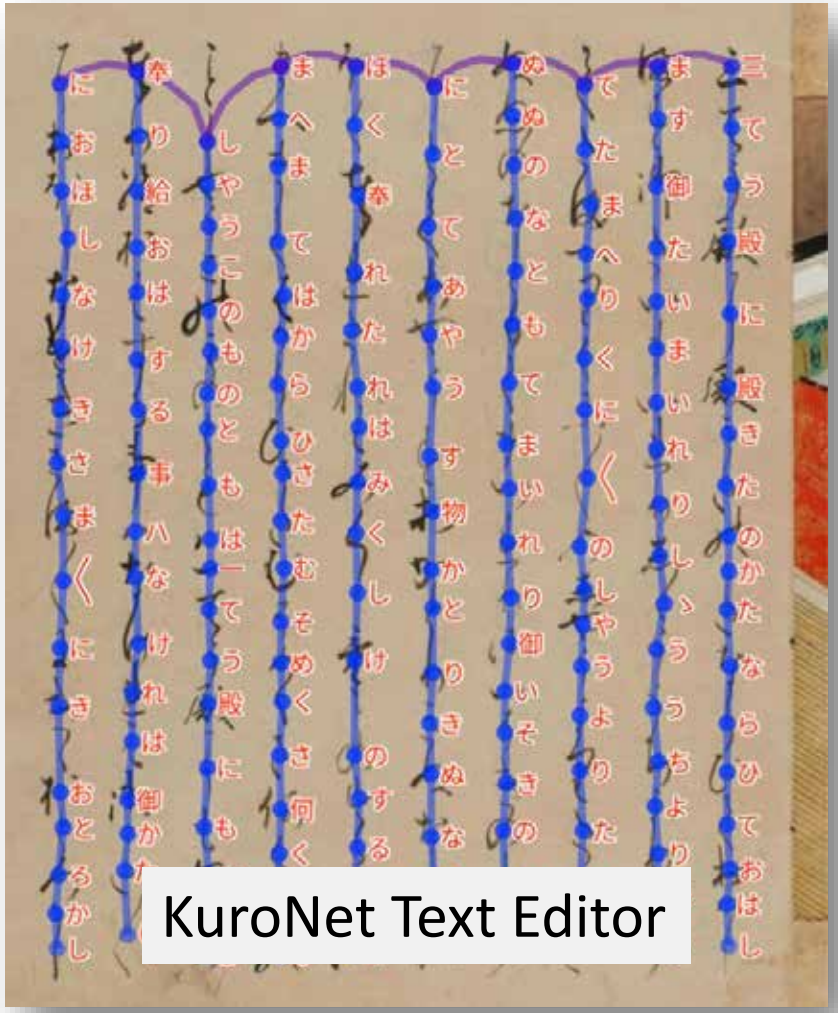
IIIF画像切り取り機能



くずし字OCR機能

KuroNetくずし字認識サービス

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



テキスト化

三てう殿に殿きたのかたならひておはし
ます御たいまいれりしゝううちよりまう
てたまへりくにくのしやうよりたうき
ぬぬのなともてまいれり御いそきのれう
にとてあやうす物かとりきぬなとお
ほく奉れたれはみくしけのする人御
まへまてはからひさたむそめくさ何くれの
しやうこのものともはーてう殿にもわかち
奉り給おはする事ハなけれは御かたく
におほしなけきさまくにきおとろかし

テキスト化サービス

kaggle くずし字認識コンペ

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



機械学習エンジニアが300万人以上登録する、世界最大のデータサイエンスプラットフォームKaggleにて、くずし字認識コンペを開催。

- 期間：2019年7月19日～10月14日
- 参加チーム数：293
- 参加者数：338
- 結果提出回数：2652

kaggleコンペの結果

上位の精度は約95%

1. さまざまな国の機械学習エンジニア、研究者が上位入賞。
2. 日本語やくずし字の知識がなくても開発可能。
3. コンペの準備を滞りなく進めるには、情報学者と人文学者の協働が不可欠。

#	△pub	Team Name	Notebook	Team Members	Score	Entries
1	—	tascj			0.950	13
2	—	Konstantin Lopuhin			0.950	60
3	—	Kenji			0.944	161
4	▲1	YoudaoOCR			0.942	49
5	▼1	See--			0.940	42
6	—	abc			0.939	15
7	—	K_mat			0.934	20
8	—	t-hanya			0.920	21
9	—	Ollie, Nanashi, and Tom			0.910	35
10	—	Zenkei_R&D			0.903	144
11	—	masayai			0.903	12
12	▲5	Kirill Brodt (shad nsk)			0.901	4
13	▲1	James Day			0.901	33
14	▼1	NEU			0.900	54
15	▼3	s tatsuya			0.900	29

AIくずし字認識

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

日本古典籍データセット（国文研蔵）



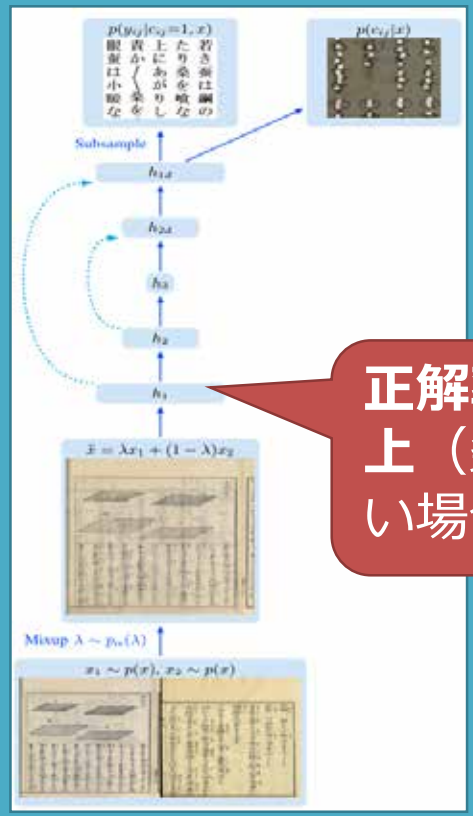
くずし字データセット（国文研・CODH作成）



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



くずし字認識モバイルアプリ

江戸マップ – 江戸切絵図の地名リソース化

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



番号	分類	現代語訳	翻刻	地図
2-001	施設	幸橋御門	幸橋御門	拡大図
2-002	施設	山下御門	山下御門	拡大図
2-003	施設	数寄屋橋御門	数寄屋橋御門	拡大図
2-004	施設	鍛冶橋御門	鍛冶橋御門	拡大図
2-005	施設	呉服橋御門	呉服橋御門	拡大図
2-006	地名	一石橋	一石橋	拡大図
2-007	地名	出橋	出橋	拡大図
2-008	町名	丸屋町	丸屋丁	拡大図

29枚の地図から8658ヶ所の地名を抽出してデータベース化

江戸切絵図の構造化

1. 尾張屋版（1849-1862）＝江戸切絵図の中で最も大量に売れて普及した地図。うち29枚を対象とする。
2. **国立国会図書館「デジタルコレクション」**で公開されるIIIF画像を、IIIF Curation Viewerで読み込み。
3. 地名を矩形で囲んで画像座標を記録し、矩形中の文字を翻刻・現代語訳し、データベース化。
4. 「施設」「屋敷地」「寺社仏閣」「店名」「地名」「町名」「海川池」「観光地」「その他」の9分類を付与。ただし旗本屋敷などは数が多いため除く。

古地図のジオレファレンス



GCPの対応付け





国立国会図書館
『江戸切絵図』

立命館大学
日本版Map Warper

江戸マップβ版＋日本版Map Warperタイル配信サービス


千代田区
© 2020 ZENRIN
Google Earth

史料とエンティティのリンク



宇田川丁、三島丁・神明丁、此分潰家多、土蔵残所なし

固有表現認識

宇田川丁、**三島丁**・**神明丁**、此分潰家多、土蔵残所なし

御江戸大地震大破并出火類焼場等書上之写（みんなで翻刻）

1. 江戸マップに出現する地名を、地名リソース（エンティティのデータベース）として整備する。
2. 史料の文字列から固有表現（地名）を抽出する。
3. 特定のエンティティとリンクすることで、実世界と紐づける。

曖昧性解消

原資料表記	江戸マップID	江戸マップ表記
宇田川丁	4-358	**宇田川町**
三島丁	4-290	**三島丁**
神明丁	4-294	**神明町**

江戸買物案内 – 商業ビッグデータ

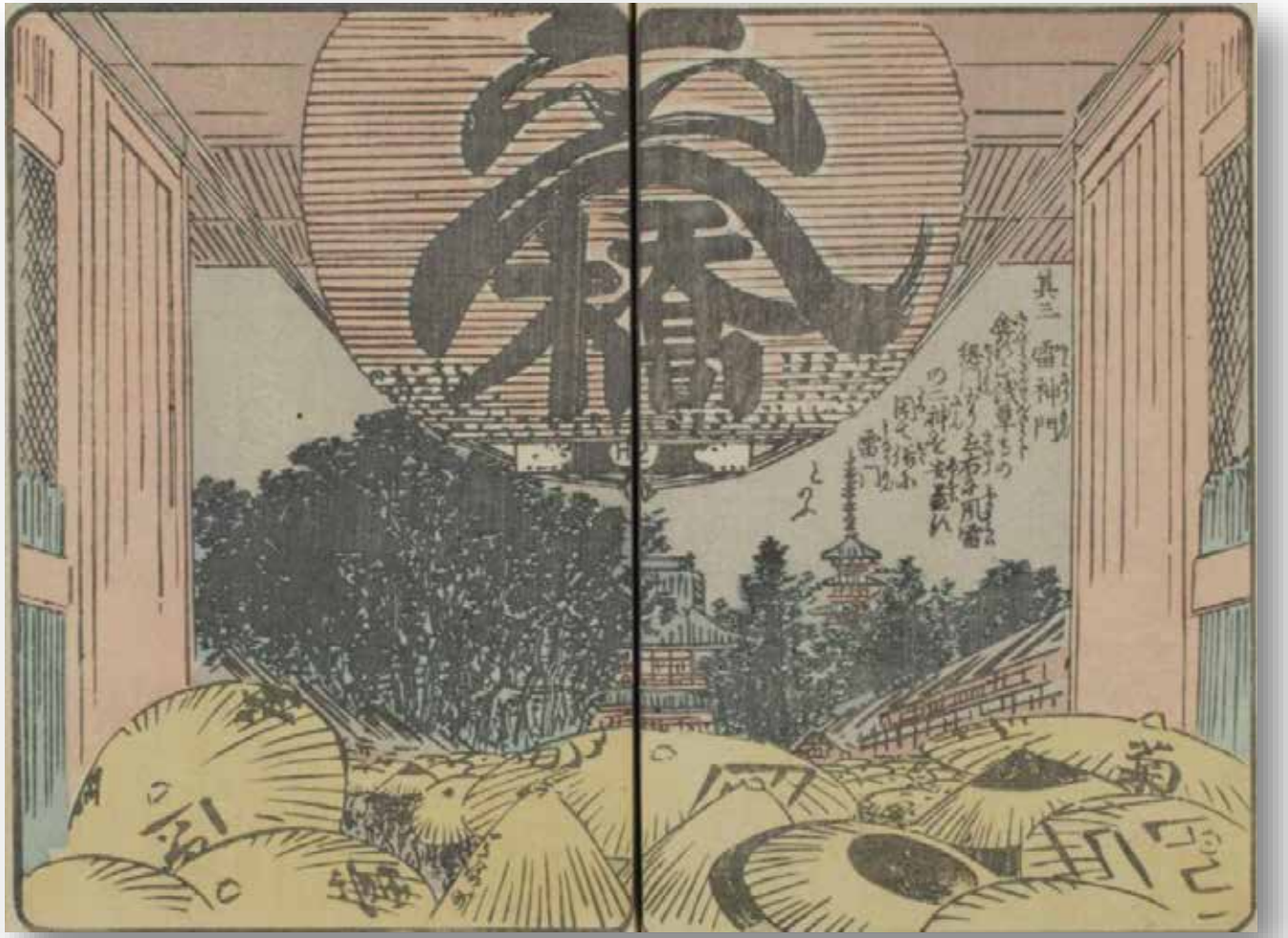
http://codh.rois.ac.jp/edo-shops/



- 『江戸買物独案内』（1824）から、商人名や居所などを抽出しデータベース化。
- 歴博の「江戸商人・職人データベース」は現在の区レベル。
- 江戸買物案内は、江戸マップの町レベルでリンク。

江戸観光案内 – 観光ビッグデータ

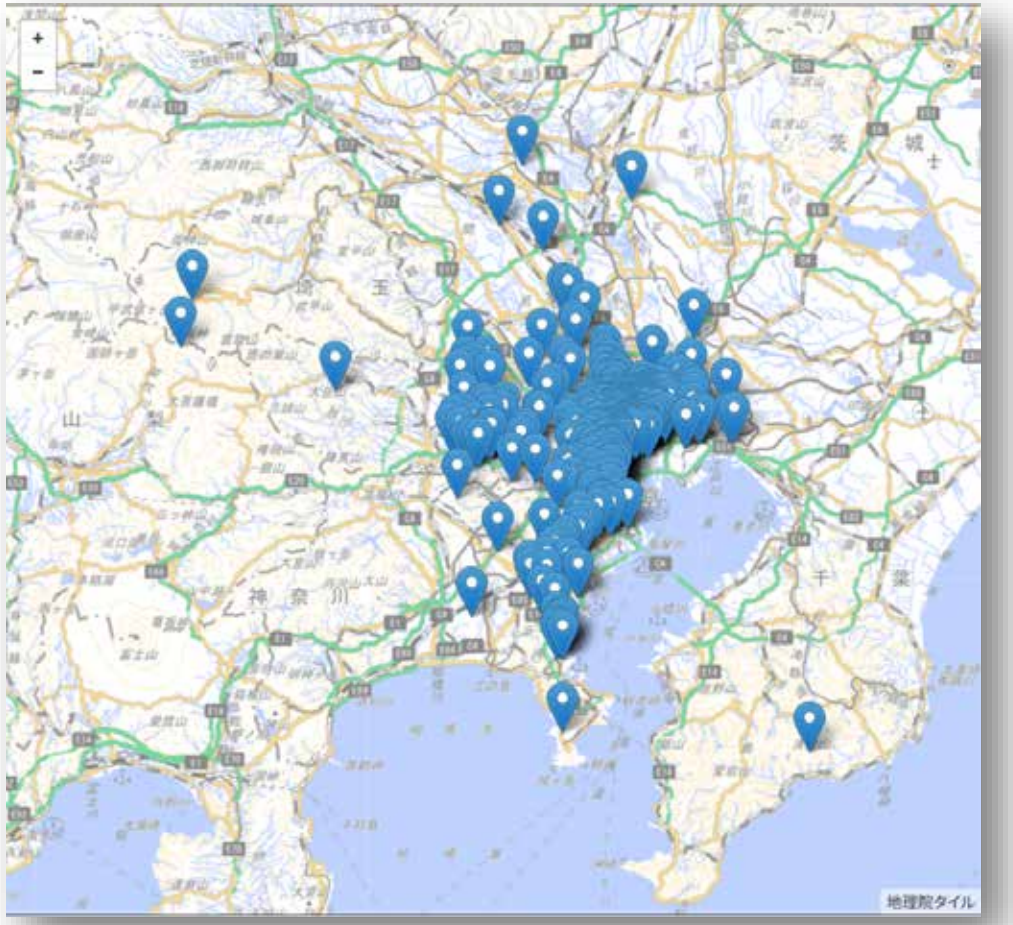
http://codh.rois.ac.jp/edo-spots/



- 江戸時代の名所記や名所案内を各世紀2点ずつ選択。
- IIIF Curation Viewerを用いて挿絵を切り抜き、地名を翻刻し、メタデータを付与。
- 江戸マップ、歴史地名データとリンク。

CODH 鈴木 親彦 ほか

江戸観光案内の地名リンキング



- **1255件**の挿絵を収集。その中で地名が明確な挿絵は**1219件**。
- **歴史地名データ**とのリンクは**765件（全体の6割強）**で可能。
- 歴史地名データは明治時代の地名が多いが、観光名所の地名は江戸と明治で大きな変化はなかった可能性がある。

歴史的行動記録での活用



- 清河八郎が安政2年（1855）に江戸を訪問した日記『西遊草』を分析。
- 日記から地名を抽出し、江戸マップや江戸観光案内とリンク。
- **164件**の地名中、**江戸マップと132件**、**江戸観光案内と119件**がリンク可能。

歴史的記録

1. **歴史的状況記録**
   - 世界の状況を観察した記録（空の状況を視覚的に観察した記録、地震の揺れを聴覚・触覚で感じた記録、地震の被害状況を視覚で観察した記録など）
2. **歴史的行動記録**
   - 人間の行動に関する記録（人間がある場所から別の場所に移動した記録、店で何かを買ったり食べたりした記録、観光した記録など）
3. **歴史的状態記録**
   - 人間やモノの状態に関する記録（人間やモノの属性に関する記録、市場取引の価格に関する記録、モノの輸送に関する記録など）

れきすけ

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



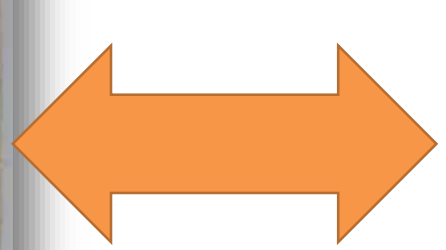
- 史資料の所在情報を共有するサービス。
- 書誌情報に加え、事項ごとの歴史的記録の所在に関する情報を共有できる点が新しい。
- 「地震のことが書いてある日記を検索したい」を実現する。

CODH 市野 美夏ほか

アナログとボーンデジタルの接続



みんなで翻刻
https://honkoku.org/?#/transcription/576333664A57AF196A194E8115FAEF13/1/





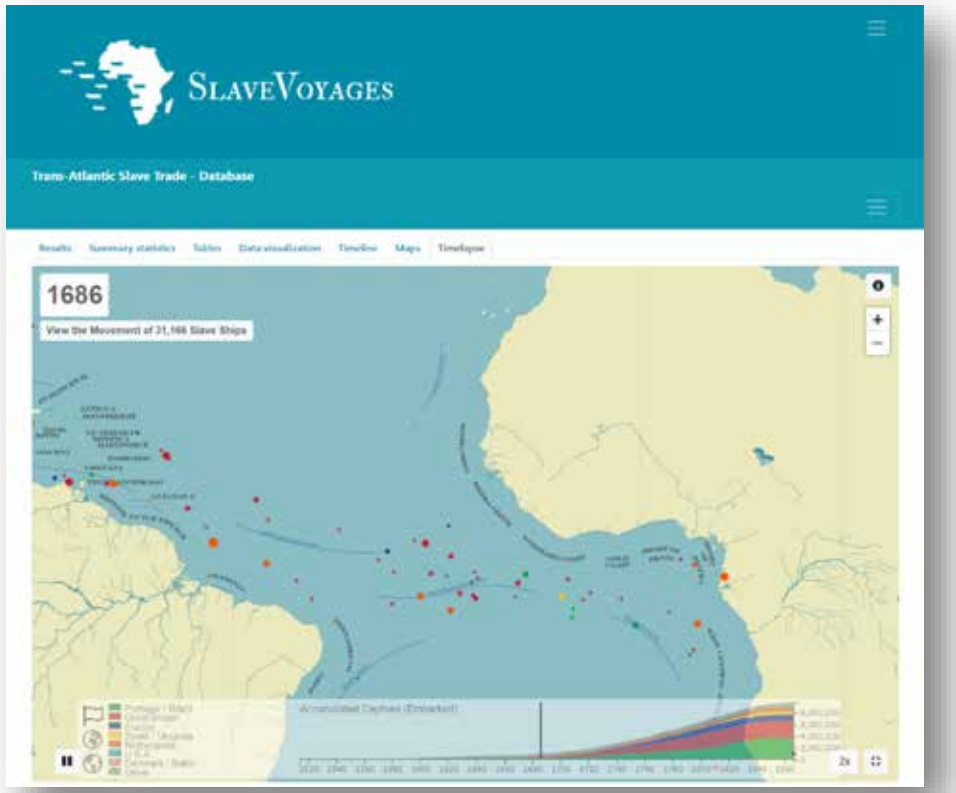
地震なう！

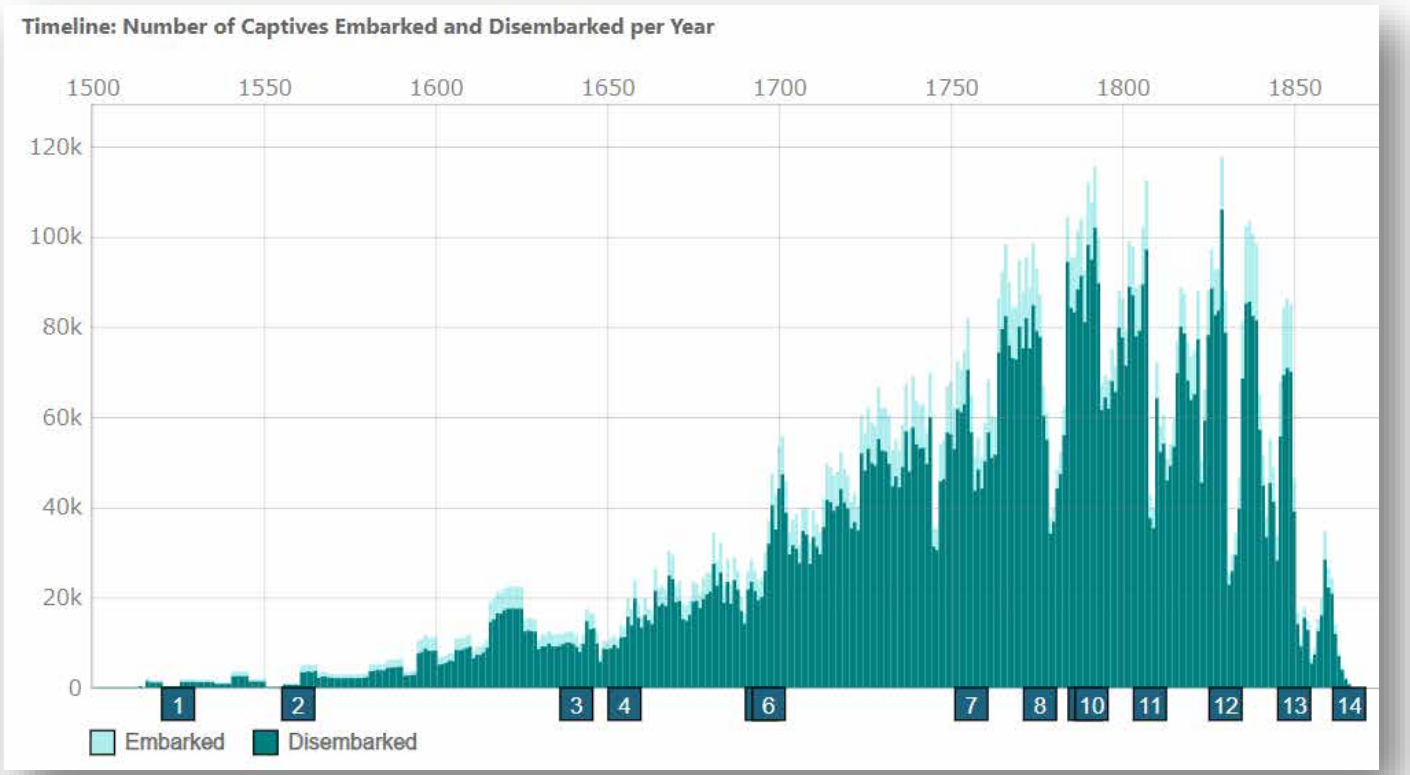
アナログをデジタル化して両者を接続し、現代の技術（プラットフォーム）を過去に延長していく。

人間と機械が共に読む時代

1. 過去の膨大な文書は、人間が読むために作られてきた。
2. これからは人間と共に機械も読み、相互に助け合う時代になる。
3. 機械可読形式への変換技術基盤が必要（ビジネス分野では**RPA**とも呼ばれる）。
4. 人間には困難な部分を機械が代替・支援し、そこから新しい価値を生み出していく。

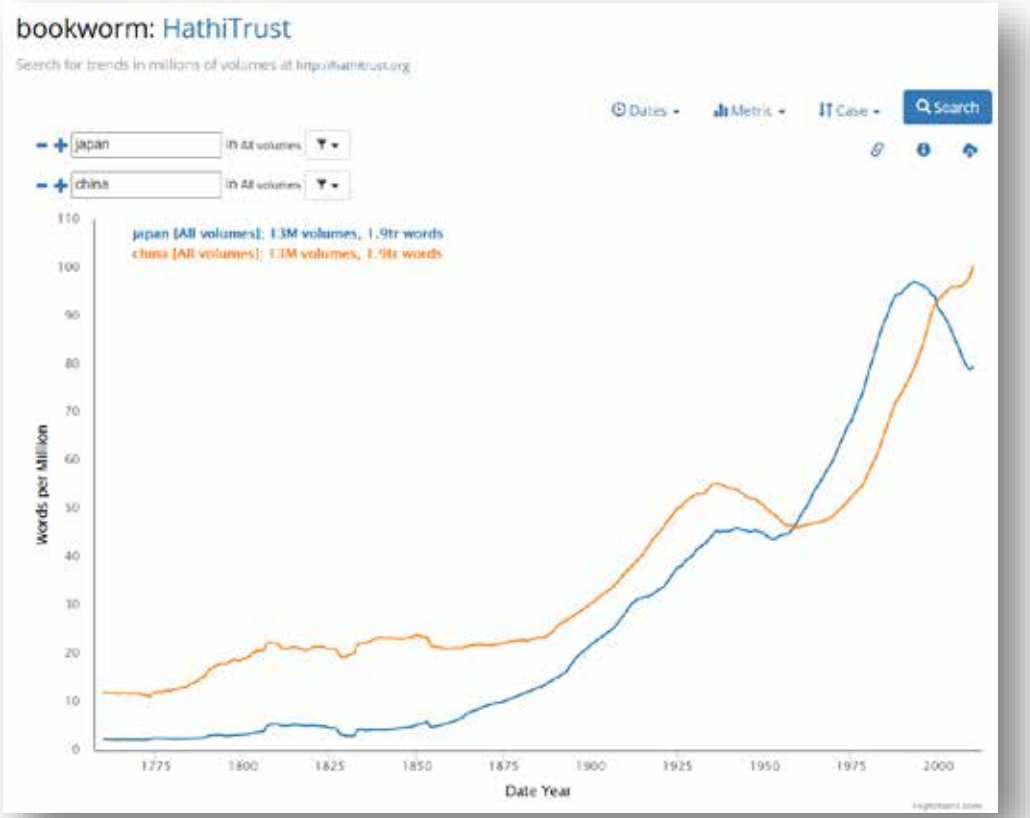
ビッグデータと可視化





Slave Voyages Trans-Atlantic Slave Trade, https://www.slavevoyages.org/

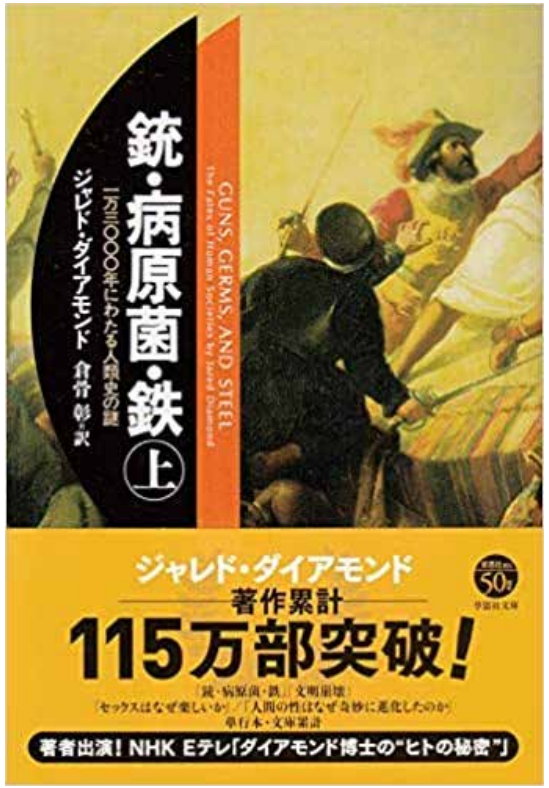
ビッグデータと検索



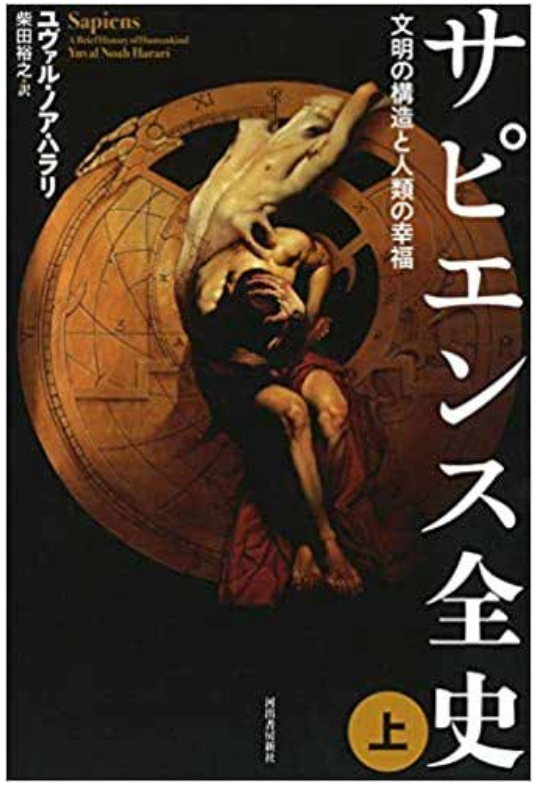
HathiTrust bookworm,
https://bookworm.htrc.illinois.edu/develop/

- 人文学ビッグデータに基づく語（N-gram）の出現頻度の推移。Google Books Ngram Viewer や HathiTrust bookworm などで検索可能。
- 検索機能が人文学研究のスピードとスタイルを変える。
- OCRが困難なため、日本語では同様の検索は困難。

グローバル／ビッグヒストリー



ジャレド・ダイアモンド、銃・病原菌・鉄



ユヴァル・ノア・ハラリ、サピエンス全史

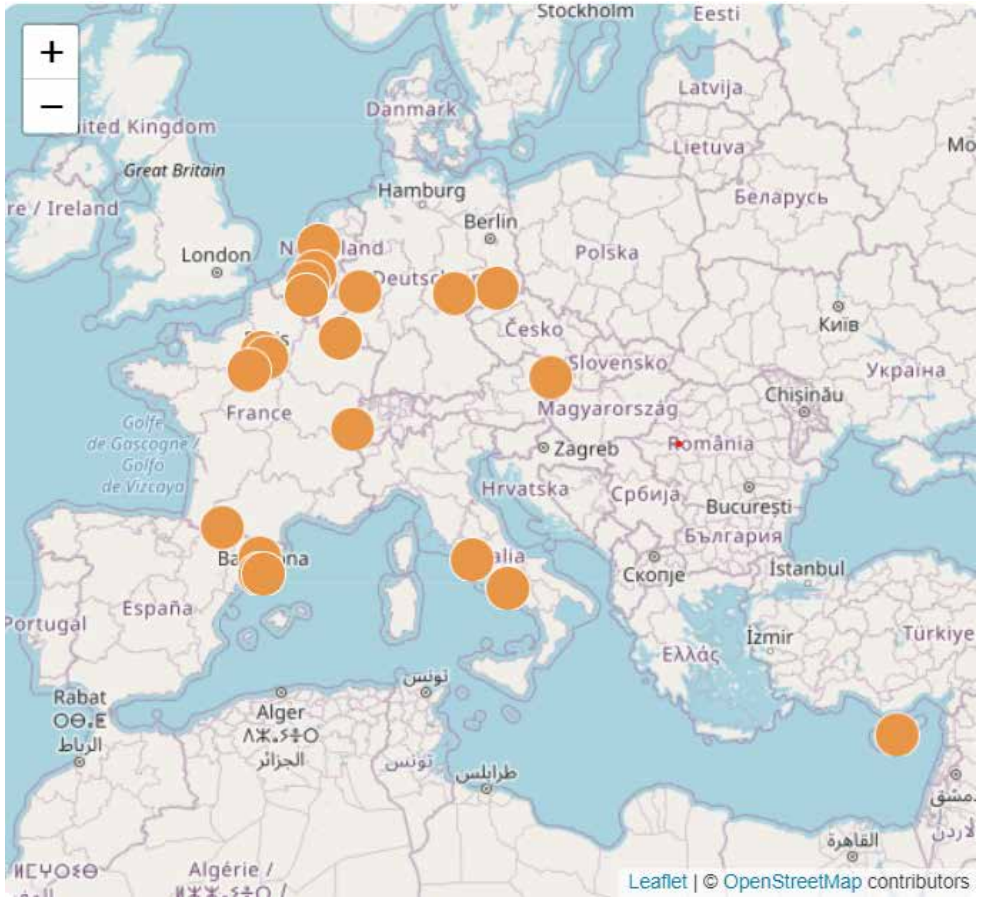


デヴィッド・クリスチャン、ビッグヒストリー われわれはどこから来て、どこへ行くのか――宇宙開闢から138億年の「人間」史

「歴史」の対象が空間的・社会的に拡大し、より分野横断的になってきた

Time Machine Europe

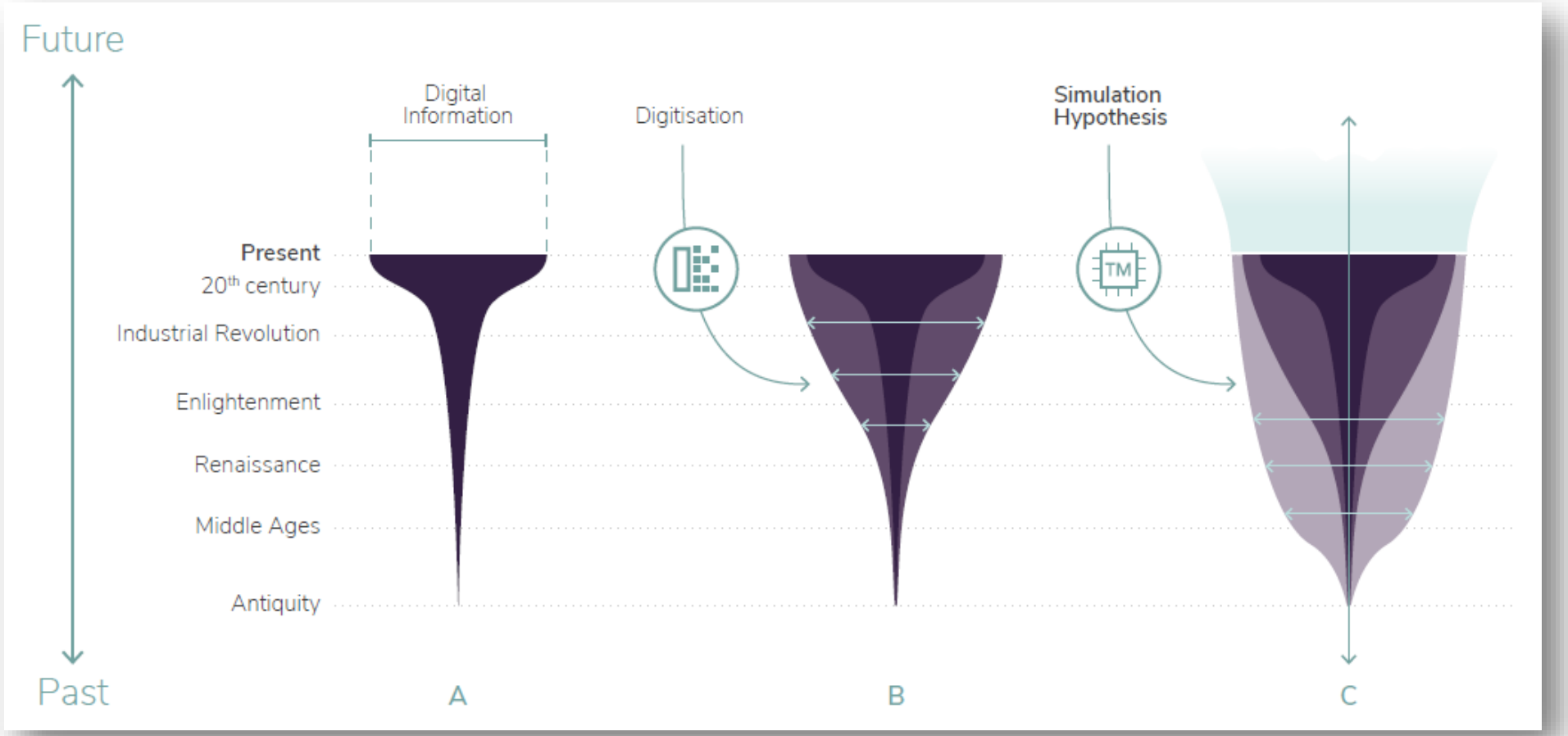
https://www.timemachine.eu/



- 欧州の文化産業や観光、教育などにも有用な、誰もがアクセスできる「過去のビッグデータ（Big Data the Past）」を構築。
- 革新的なデジタル化技術や人工知能（AI）技術などを適用。
- 参考：「過去のビッグデータ」を探る欧州タイムマシン研究計画、カレントアウェアネス-E、https://current.ndl.go.jp/e2248

LTM Projects, https://www.timemachine.eu/ltm-projects/

4D Mirror World



Time Machine Manifesto, https://www.timemachine.eu/wp-content/uploads/2019/06/Time-Machine-Manifesto.pdf

歴史のビッグデータ化

1. 過去の歴史の「**記述の断片（例：5W1H）＝記録**」を構造化し、機械可読化し、データベース化する。
2. 歴史的記録を数値化し、様々な観点から可視化することで、データの傾向を大まかに把握する。
3. 歴史的記録をボトムアップに組織化し、人間が理解しやすいストーリー（物語）化する。
4. 歴史をモデル化し、シミュレーションやシナリオ予測することで、現在や未来に資する知識を生み出す。

謝辞

- 本発表の内容には、ROIS-DS人文学オープンデータ共同利用センターの鈴木 親彦、市野 美夏、カラーヌワット タリン（特任助教）の成果が含まれています。
- 本研究の一部は、科学研究費「歴史ビッグデータ研究基盤による過去世界のデータ駆動型復元と統合解析, 基盤研究(A), 19H01141, 2019-2021」の支援を受けました。



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

Twitter: @rois_codh

Facebook, YouTube, GitHub