資料２－２
科学技術・学術審議会　学術分科会
人文学・社会科学特別委員会（第７回）
令和３年６月２８日

人文学におけるデータ駆動型研究の類型と事例

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
北本 朝展（KITAMOTO Asanobu）
https://researchmap.jp/kitamoto/

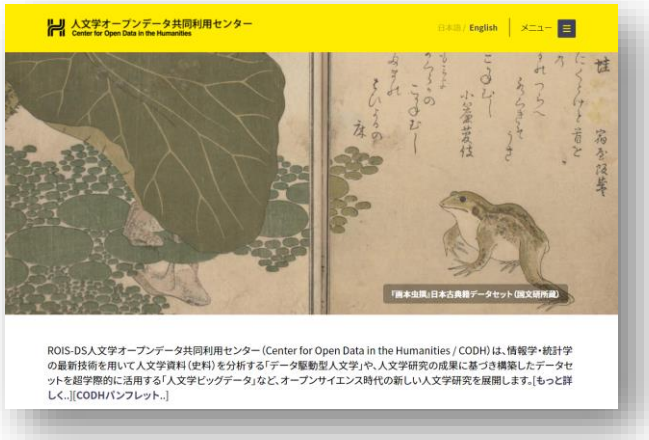
自己紹介

https://researchmap.jp/kitamoto/

- **北本朝展（きたもと あさのぶ）**
- 国立情報学研究所 教授
- 2016年 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
- 1997年 東京大学大学院 工学系研究科 電子工学専攻修了／博士（工学）
- 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研究コラボレーションにも興味を持つ。

ROIS-DS人文学オープンデータ共同利用センター（CODH）

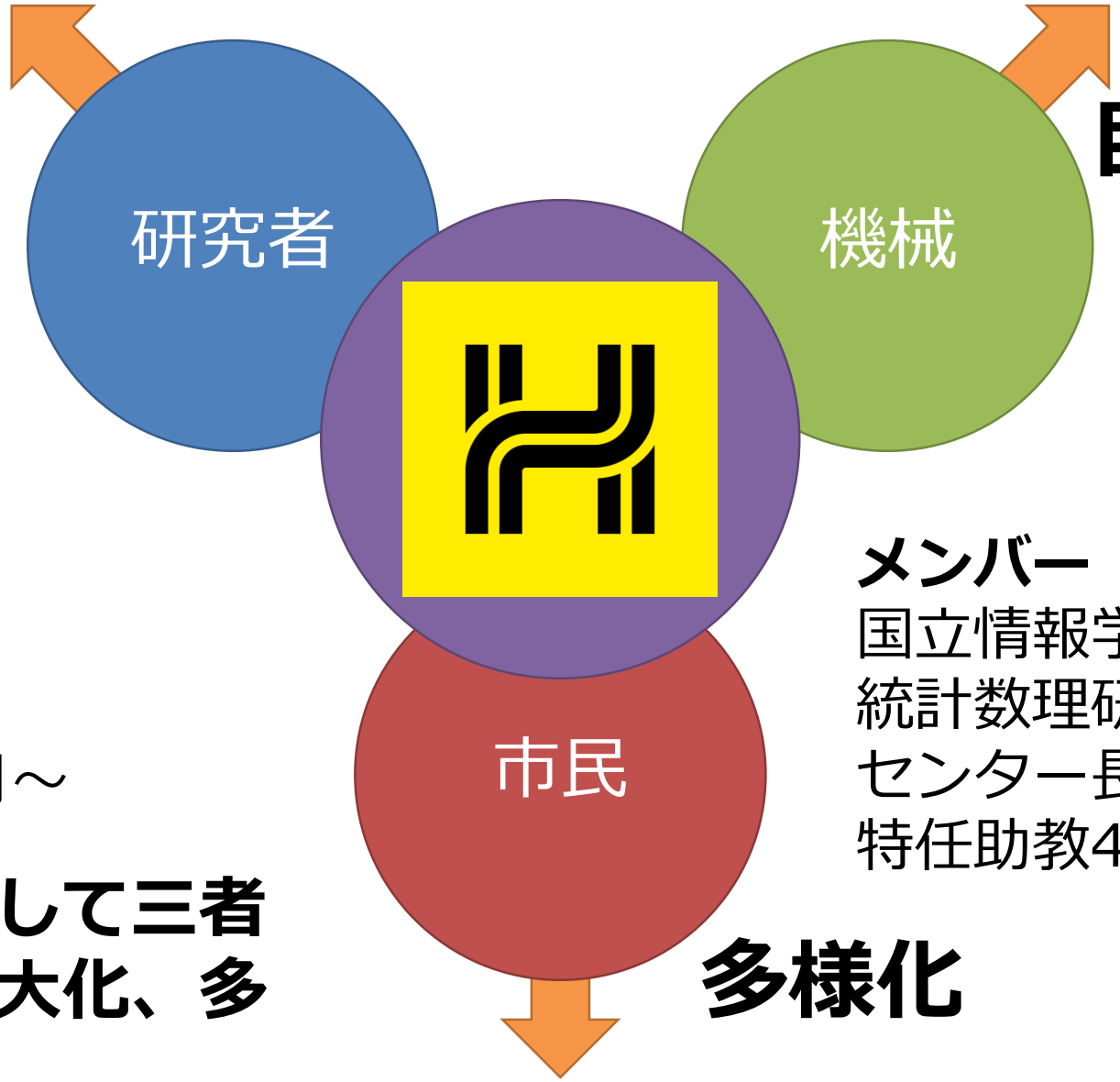
http://codh.rois.ac.jp/



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

人文学における研究データ

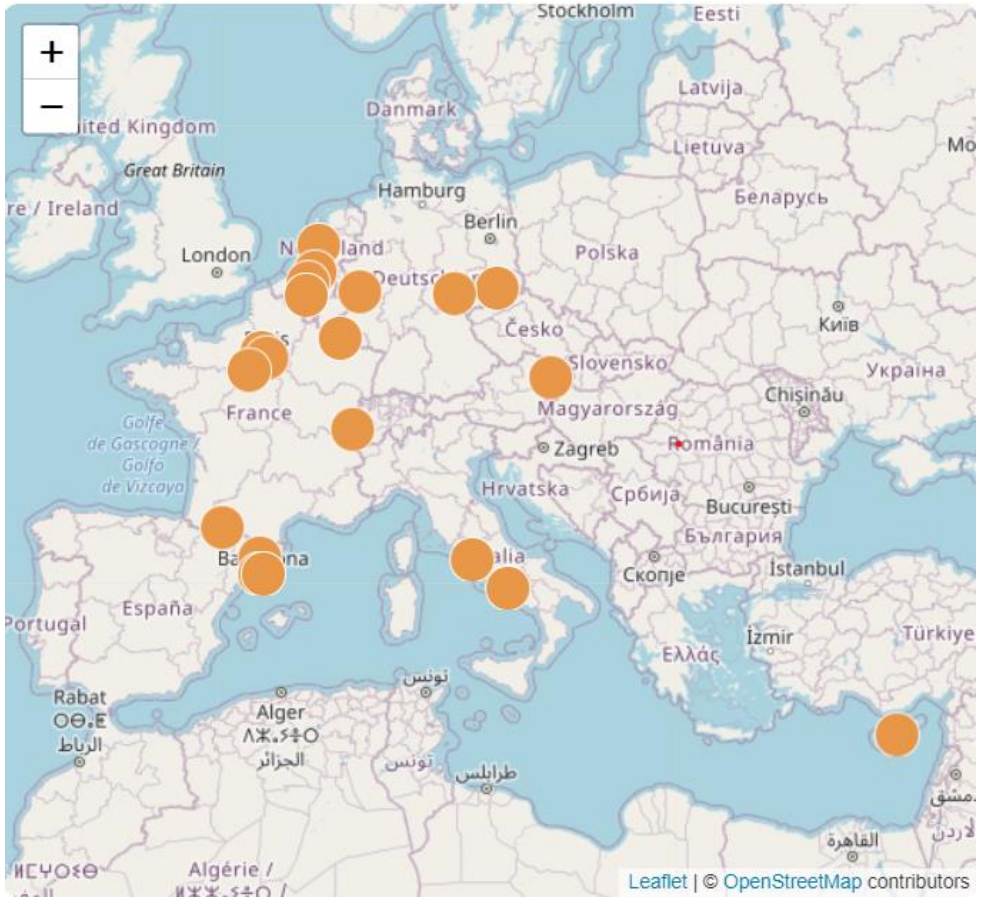
- **研究資源データ（研究の入力となるデータ・参照されるデータ）**
  - テキストデータ（人文学データの主力）
  - 画像データ（デジタル化の出発点、テキスト化できない情報もある）
  - 写真・映像データ（現地調査などでは主力）
  - 3Dデータ（立体物、遺跡、景観など）
  - 典拠データ（識別子およびメタデータ）
- **研究成果データ（研究の出力となるデータ）**
  - 論文（書籍）付属データ
  - 注釈データ（付加価値情報）
- **研究過程データ（研究の入力と出力の間にあるデータ）**
  - （半）構造化データ（研究成果データにもなる）
  - 翻刻、注釈、翻訳データ（研究成果データにもなる）

海外の動向

1. 欧州：**DARIAH**（デジタル人文学）、**CLARIN**（言語資源）＝研究基盤として着実に発展。
2. 欧州：**TextGrid**など、分野ごとに生み出される研究基礎データ（TEI）のリポジトリも拡大。
3. **欧州タイムマシン研究計画**：「過去のビッグデータ」研究に加え、文化遺産機関とも幅広く連携。教育・観光・創造産業などへの波及効果も重視。
4. 米国：**HathiTrust**（書籍デジタル化）など。

Time Machine Europe

https://www.timemachine.eu/



- 欧州の文化的資産をデジタル化・構造化し、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）技術などを開発。
- 欧州委員会のフラグシップ研究計画＝10年で1000億円（見直し中）。参加機関700以上。
  https://current.ndl.go.jp/e2248

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

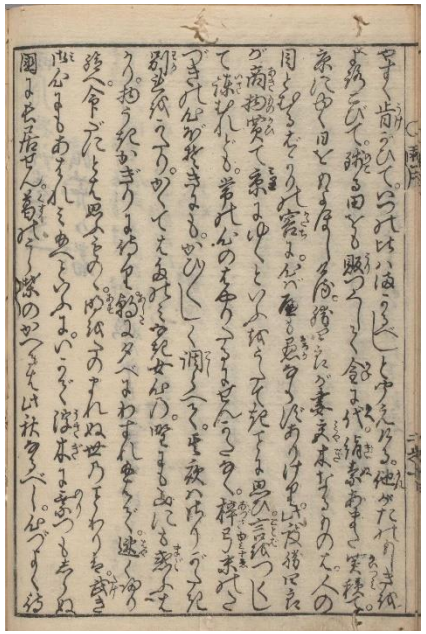
データ駆動型人文学研究の類型

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として、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。
4. 専門人材を含むチーム構成：ソフトウェアエンジニア、データクリエイター、データキュレーター、デザイナーなどとの協働により、データ駆動型人文学研究の質と量を向上させる。

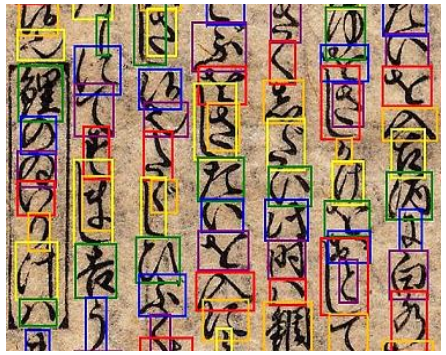
くずし字データセットの利活用例

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

日本古典籍データセット（国文研蔵）



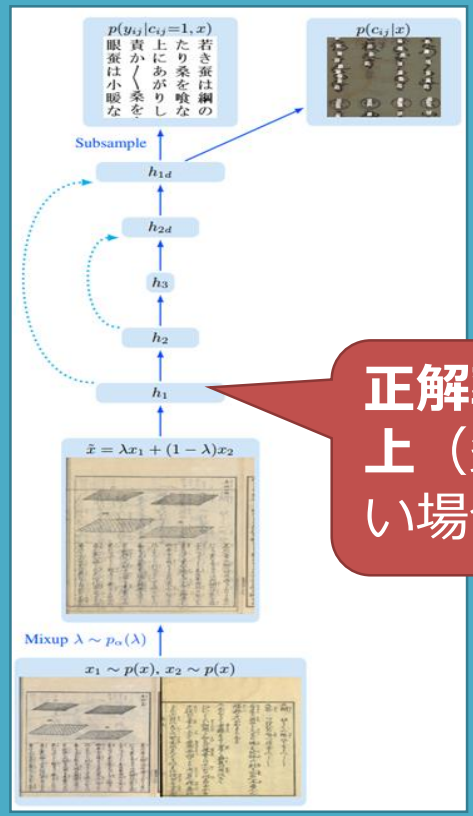
くずし字データセット（国文研・CODH作成）



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

くずし字認識モバイルアプリ

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。

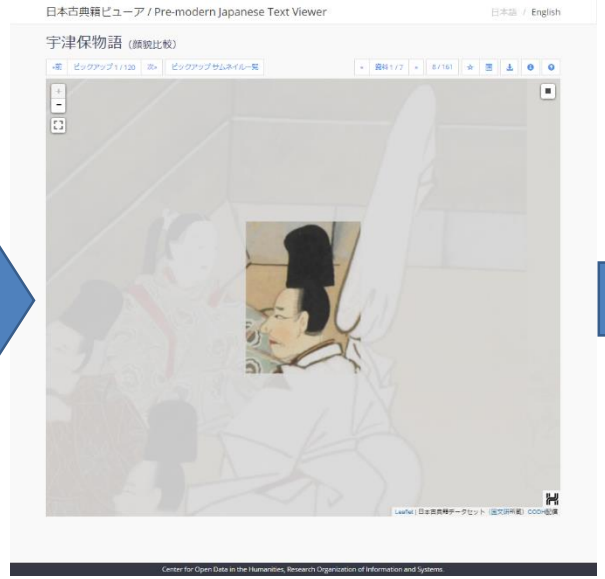
教訓：情報学者がデータの仕様策定段階から入るべき。

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

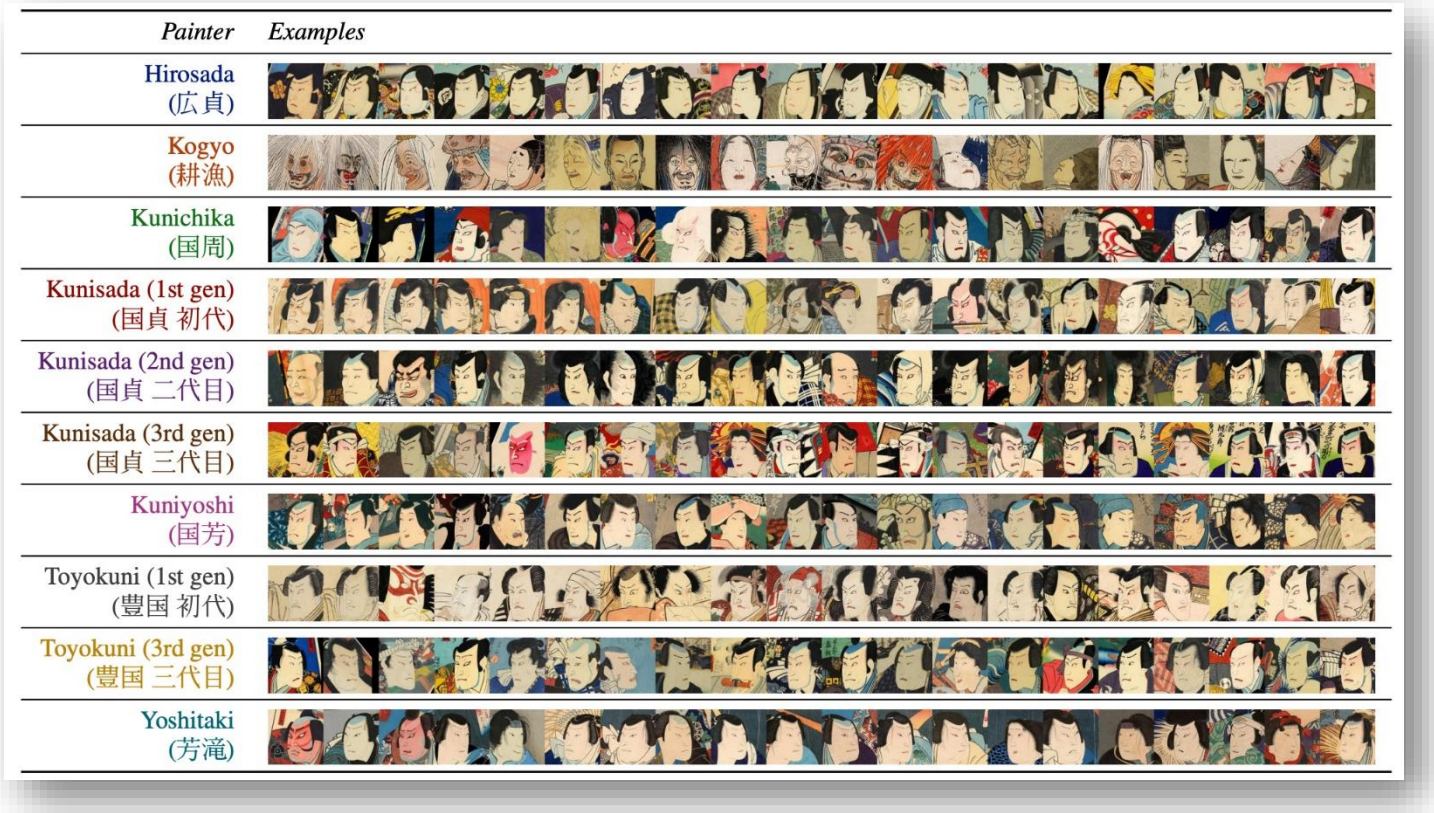
http://codh.rois.ac.jp/face/

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。

浮世絵顔データセット

http://codh.rois.ac.jp/ukiyo-e/face-dataset/



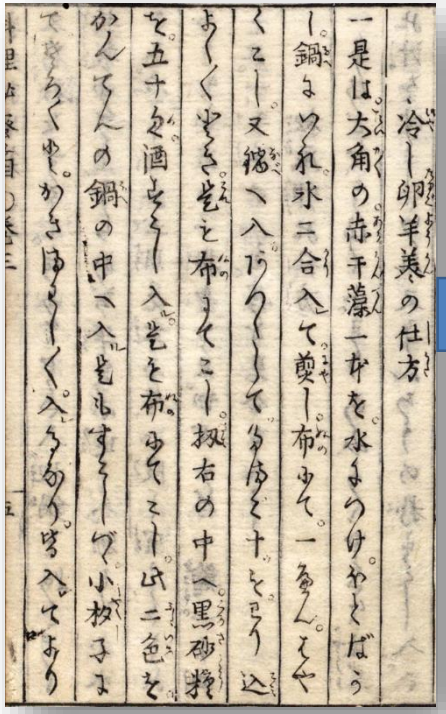
『ARC浮世絵顔データセット』（Yingtao Tian、 ROIS-DS CODH作成、 ARCから収集）, https://doi.org/10.20676/00000394

1. NII-IDRが立命館ARCと協力して浮世絵データを公開。
2. Googleの機械学習研究者が既存のAPIを使って顔を切り抜けることを発見。
3. 浮世絵の画像解析研究を発展させ、データセットを公開。

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/

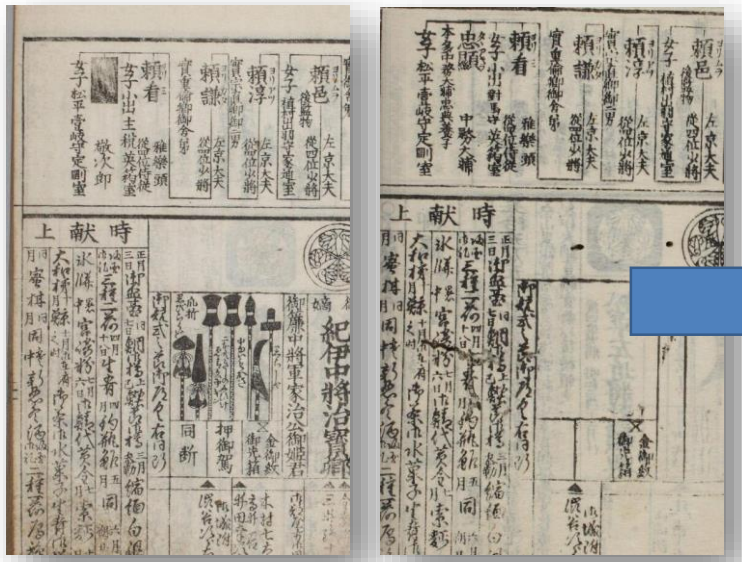
1. 国文研が主催するアイデアソンに、情報学者である北本が参加して、江戸の料理本に触れる。
2. レシピに写真を加え、料理レシピサイトに投稿すれば、現代の生活と接続できるのではないかと思いついた。
3. クックパッドに連絡して共同研究体制を構築。翻刻とレシピ化が可能なデータクリエーターに作業依頼。
4. プロの料理研究家が参画し、調理可能なレシピ化と写真撮影を進め、現代にも通用する水準のレシピを完成。

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。

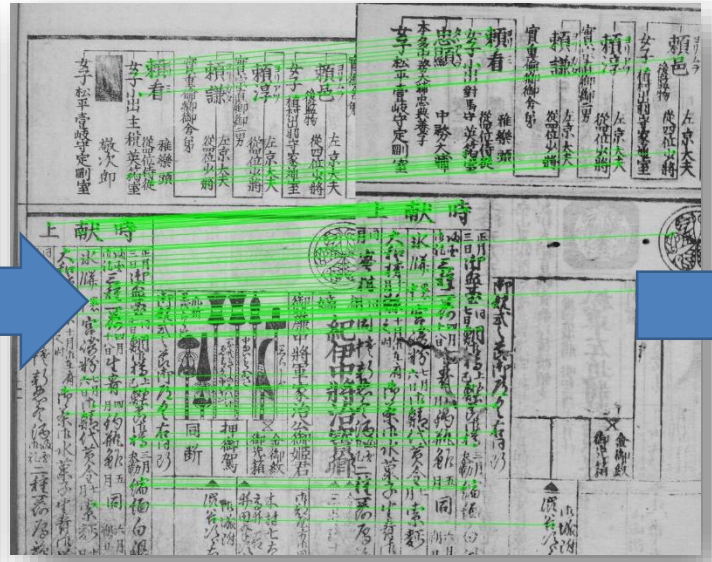
武鑑全集のワークフロー

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

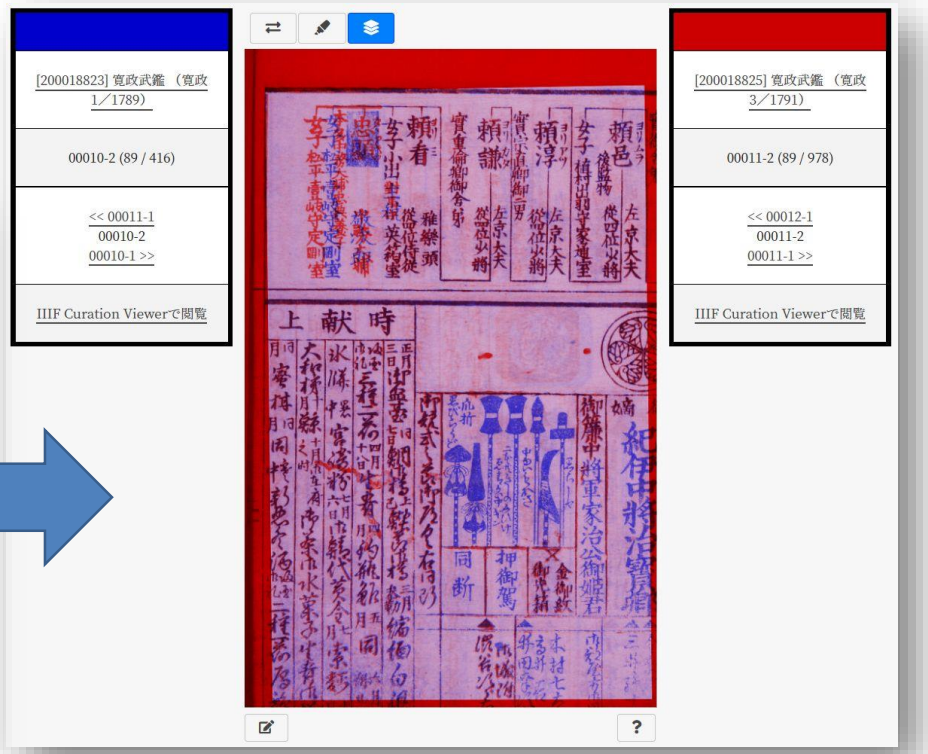
多数の版を視覚的に比較して差分を取り出す



江戸時代に頻繁に更新され出版された、大名・幕府役人に関するデータブック（国文研蔵）



コンピュータビジョン技術を用いた2枚の画像のマッチング



版間の差分を強調した「差読」の実現

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. この技術は15年以上前から存在したが、それが木版印刷本の版間差分の強調に使えるという発想はなかった。
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。

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

人文学者のプロジェクトに対して、情報学者の立場から、データセットの構築に協力。

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。

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

自然科学的データ

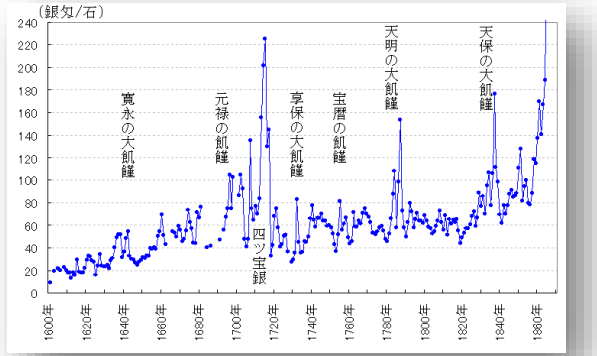
- 気候
- 地震
- 噴火
- 疫病

人文社会的データ

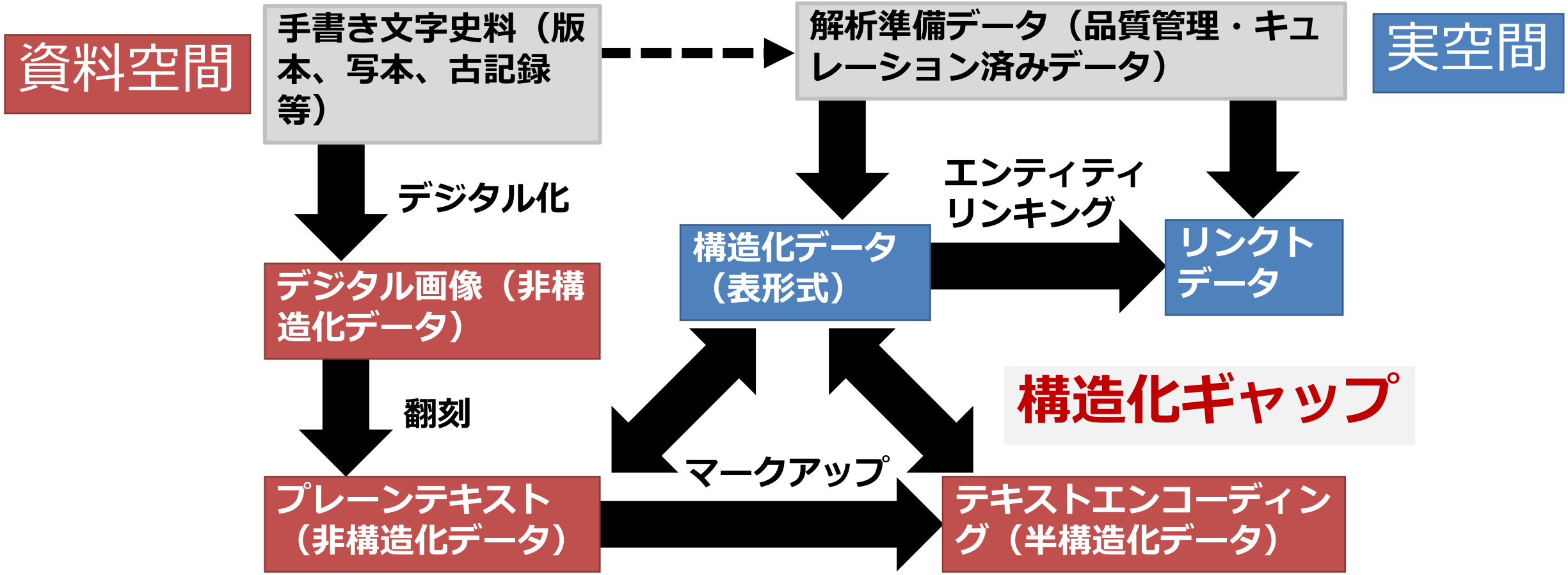
- 経済
- 人口
- 政治
- 文化

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

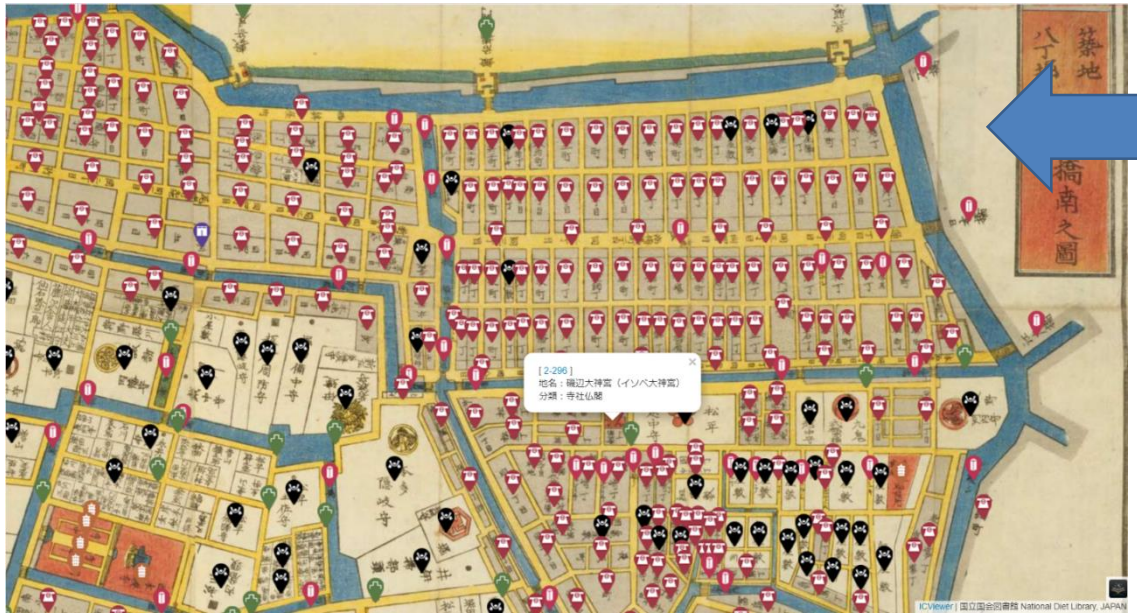


江戸ビッグデータの統合

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

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/edo+150/

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）を加えて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展開し、基盤データに独自の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。

識別子＝研究インフラ

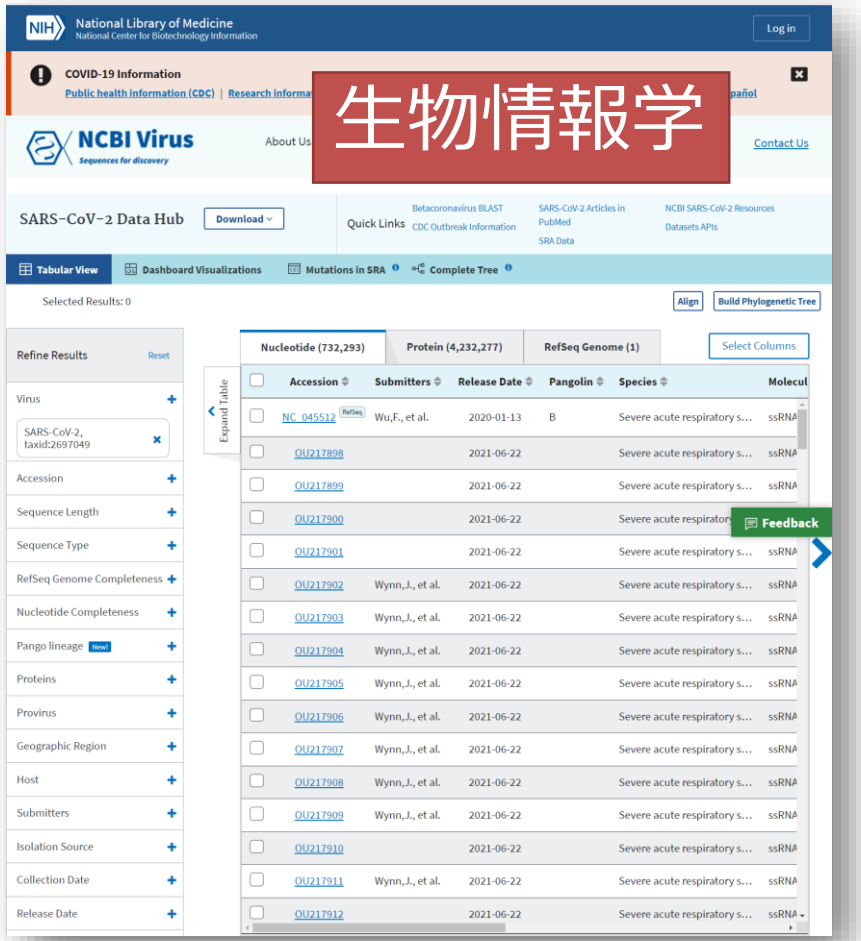
デジタルオブジェクト

https://doi.org/

研究者

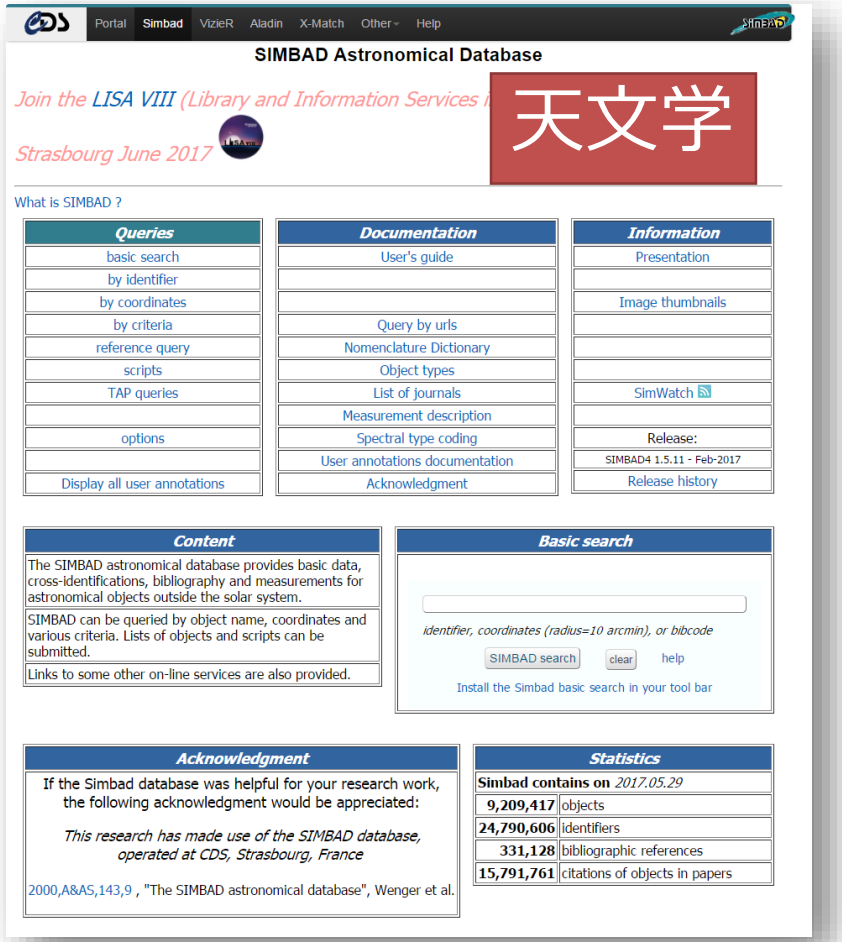
https://orcid.org/

生物情報学



https://www.ncbi.nlm.nih.gov/

天文学



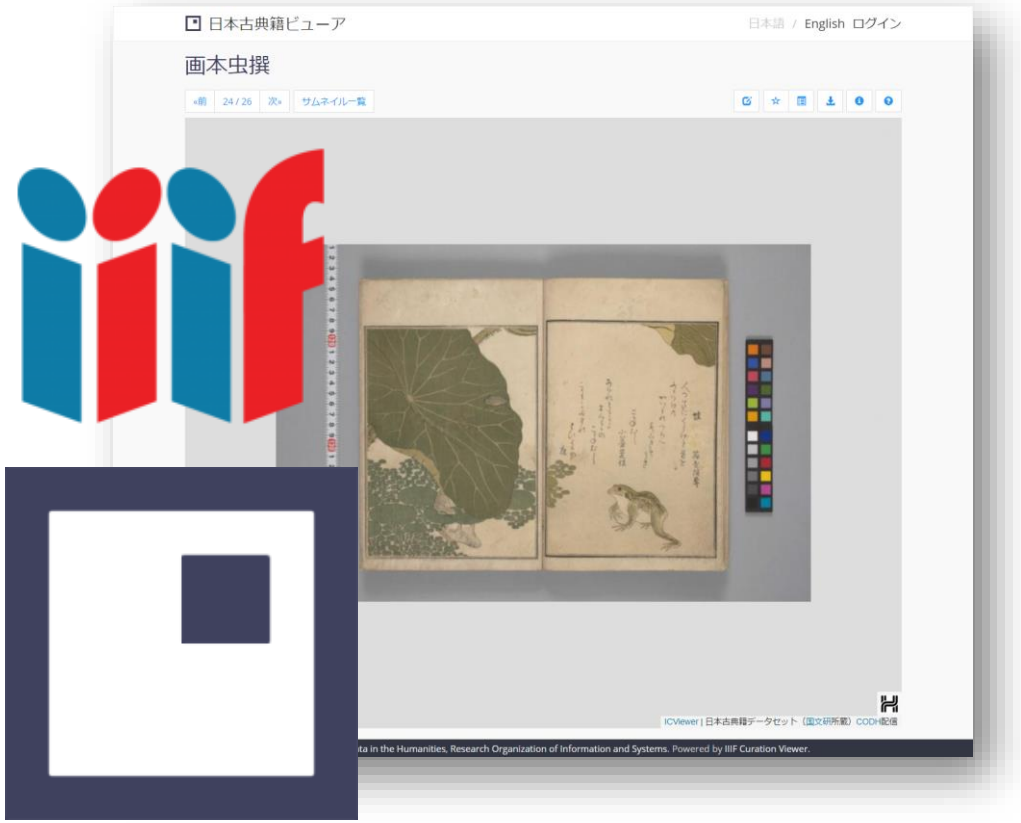
http://simbad.u-strasbg.fr/simbad/

データセットと識別子

1. データセットを構築する際に、地名、人名、時間などの固有名や分類体系（コード）などは、可能な限り共通化すべき。
2. レファレンス的な研究資源構築は、コミュニティに対するインパクトが大きいため、個別のデータセット構築とは切り分け、最初からその存在を想定した進め方がよい。
3. 権威性や専門家ネットワークなども考慮しながら、コミュニティを代表する組織や研究グループが取りまとめて進める方がよいのではないか。
4. 国際的な活動と連携する場合には、日本の代表としての意見集約や意見表明が求められることもある。

オープンソースソフトウェアの開発

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



1. 人文学データに使えるオープンソースソフトウェアを、相互運用性を尊重しながら構築する。
2. ソフトウェアエンジニアとの協働体制を確立し、高品質なコードとドキュメントを生成する。
3. 「ノーコード」の時代はシステムを作らないことも賢い選択。既存の成果をうまく活用することが、素早いDXにつながる。

人文学データ活動の拠点

1. データセットを構築する方法の細部については、分野や目的ごとに様々に異なる。
2. 共通的な概念や方法論に関する情報共有や教育などについては、それを取りまとめる機能も必要である。
3. DARIAHにおいては、marketplaceの構築、教育へのアクセス、組織の調整、政策へのアドボカシーなどがミッションとして取り上げられている。しかしこれは**開始後15年の段階**。
4. 日本は初期段階のため、基礎的な活動に重点を置き、多様性を確保し、多くの分野で、様々な可能性を試す方がよいのではないか。
5. ただし**識別子などの基盤データは乱立よりも統一**が望ましい。

新しいデータ文化の浸透

1. データに関する成果は、機械可読データとして、再利用可能なライセンスを付し、ウェブ公開することを基本とする。
2. データに関する成果の公開を、きちんと業績として評価すべき。**評価システムのアップデートは文化の課題。**
3. 従来の人文学分野は、出版社がデータに関する業務をサポート、著作物として流通させていた面が大きい。**この文化が根本的に変化することのインパクト**を共有する。
4. 人文学研究の中には、データ文化になじみやすいテーマもある。機械可読性のエッセンスさえつかめれば、情報系以上のデータ専門家になれる可能性がある。

個人単位からチーム単位へ

1. 人文学の研究は個人単位の研究が伝統的に多く、単独で進める研究の方が価値が高いという考え方も見られる。
2. 個人がすべてを行う研究体制（ワンオペ）では多種のスキルを学ぶのは難しく、方法論を進化させる余裕が乏しくなる。
3. チームで作業を分担し、協働して研究を進める文化に変えていくことが、この状況を変えるきっかけとならないか？
4. 地域・時間・作品などの軸で分担する共同研究の場合、どの分担も作業自体は似ている。データ収集、構造化、付加価値化などの作業で分担し協働するモデルも必要ではないか？

期待

人文学研究においてデータ駆動型研究を推進することで、以下のような効果が期待できる（仮説）。

1. 大規模データから新たな知識を得る機会を作る。
2. 人文学研究を円滑に推進する基盤を構築する。
3. チーム型研究に基づく研究文化を開拓する。
4. よりオープンな人文学研究を促進する。