

人文学データのオープン化 を開拓する超学際的データ プラットフォームの構築

北本 朝展（国立情報学研究所 / 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) 準備室）

山本 和明（国文学研究資料館）

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

Twitter: @rois_codh

はじめに

人文学オープンデータ共同利 用センター（CODH）

- 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が2016年4月1日に準備室を開設。2017年4月1日にセンター発足予定。 <http://codh.rois.ac.jp/>
- 1. **情報学・統計学の技術を用いて人文学の研究を行う。**
- 2. 人文学のデータを用いて情報学・統計学の研究を行う。

CODH / NII / 国文研の共同研究

人文学オープンデータ
共同利用センター
人文学データの研究者
や市民による利用を促
進するオープン化拠点。

歴史的典籍NW事業
日本の歴史的典籍30
万点をデジタル化し、
国際共同研究を推進す
る大型プロジェクト。

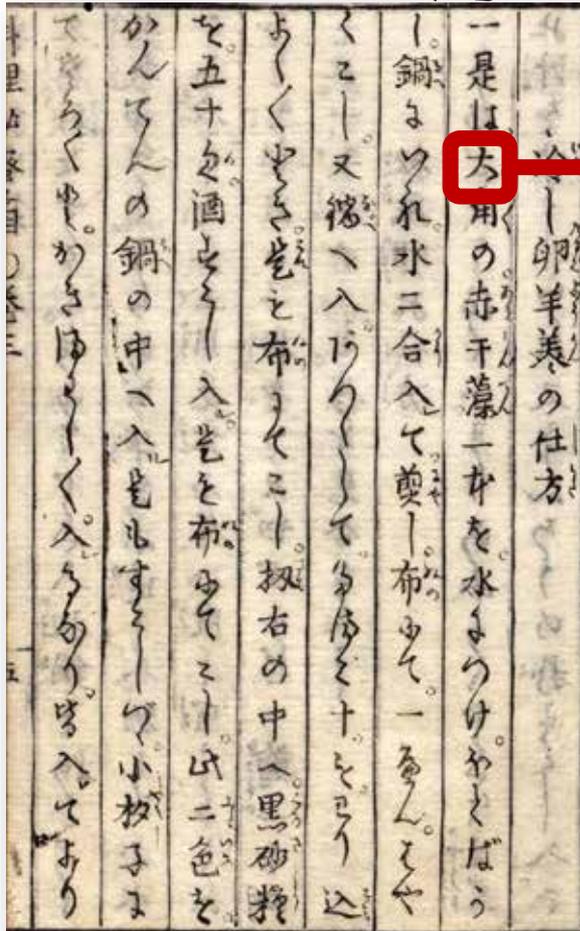
情報・システム研究機
構およびNII・統計数
理研究所が関わる。

人間・文化研究機構
国文学研究資料館が中
心的な役割を果たす。

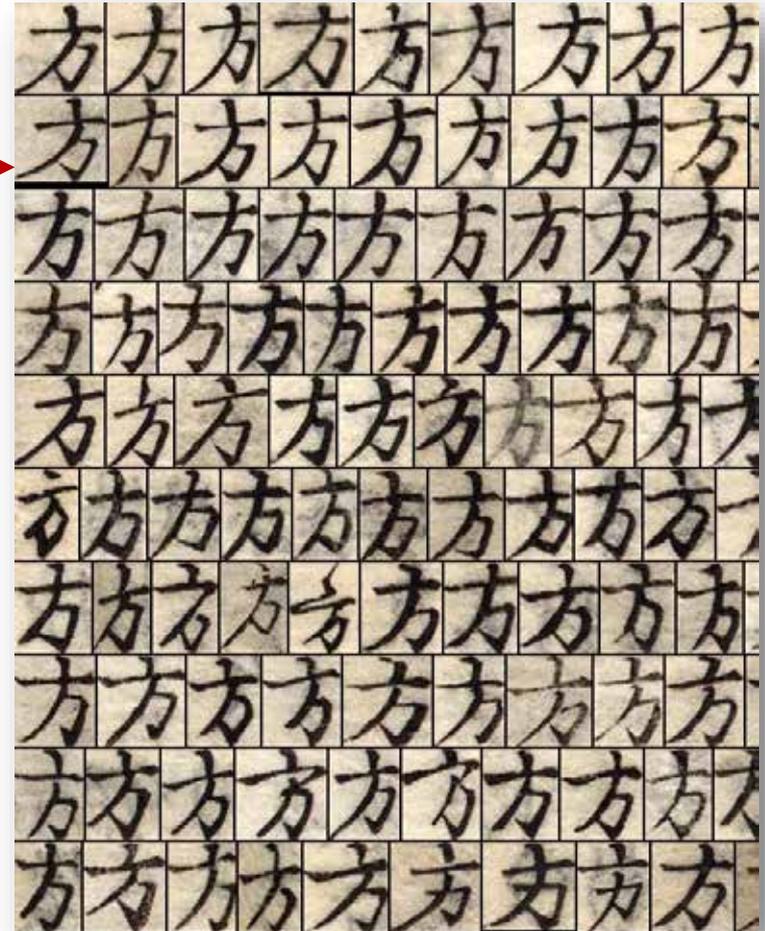
情報学と人文学の協働により歴史的典籍の課題を解決。

機械のためのオープンデータ

11月17日プレスリリース



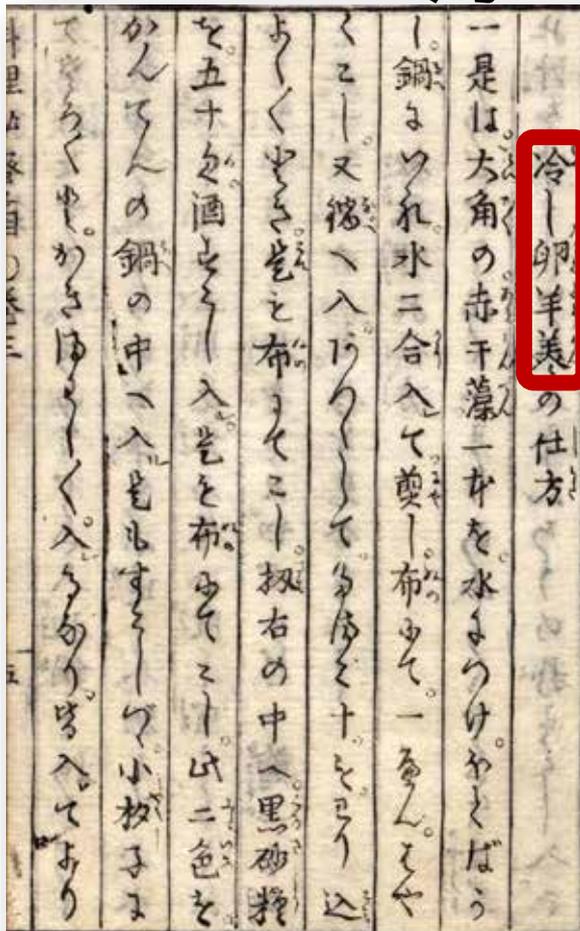
日本古典籍データセット
(国文研所蔵)



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

11月24日プレスリリース



日本古典籍データセット
(国文研所蔵)

冷し卵羊羹



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

1. 日本古典籍データ セット

<http://codh.rois.ac.jp/pmjt/>

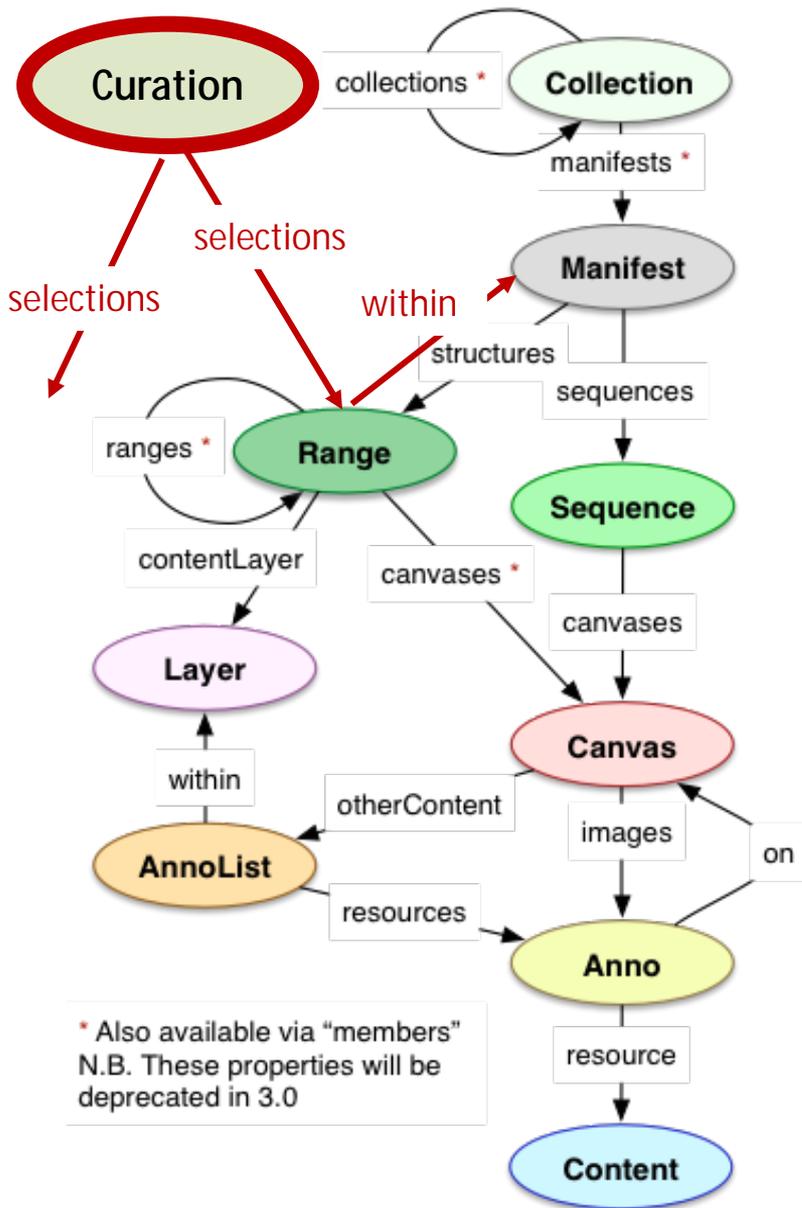
日本古典籍データセット

- 2015年11月「国文研古典籍データセット」（350点）をNIIから公開。
- 2016年11月「**日本古典籍データセット**」（**700**点）をCODHから公開。
- 画像ファイルに加えて、書誌メタデータや専門家が付与したタグデータも同梱。
- 翻刻テキストは一部の古典籍のみに付属。
- ライセンスはCC BY-SA 4.0とする。

Curation API

```
{
  "@context": [
    "http://iiif.io/api/presentation/2/context.json",
    "http://codh.rois.ac.jp/iiif/curation/1/context.json"
  ],
  "@type": "codh:Curation",
  "@id": "http://example.org/iiif/curation/curation.json",
  "label": "Curated NIJL Data set",
  "attribution": "Provided by CODH (ROIS) and NIJL NW Project.",
  "related": {
    "@id": "http://example.org/iiif/curation/sample.html",
    "format": "text/html"
  },
  "selections": [
    {
      "@id": "http://codh.rois.ac.jp/pmjt/book/200014778/range/r1",
      "@type": "sc:Range",
      "label": "Curated contents from 『画本虫撰』",
      "canvases": [
        "http://codh.rois.ac.jp/pmjt/iiif/200014778/canvas/00000",
        "http://codh.rois.ac.jp/pmjt/iiif/200014778/canvas/00011",
        "http://codh.rois.ac.jp/pmjt/iiif/200014778/canvas/00023"
      ],
      "within": "http://codh.rois.ac.jp/pmjt/book/200014778/manifest.json"
    },
    {
      "@id": "http://codh.rois.ac.jp/pmjt/book/200003067/range/r1",
      "@type": "sc:Range",
      "label": "Curated contents from 『唐糸草紙』",
      "members": [
        {
          "@id": "http://codh.rois.ac.jp/pmjt/iiif/200003067/canvas/00000",
          "@type": "sc:Canvas",
          "label": "p.1"
        },
        {
          "@id": "http://codh.rois.ac.jp/pmjt/iiif/200003067/canvas/00008",
          "@type": "sc:Canvas",
          "label": "p.9"
        },
        {
          "@id": "http://codh.rois.ac.jp/pmjt/iiif/200003067/canvas/00010",
          "@type": "sc:Canvas",
          "label": "p.11"
        }
      ],
      "within": {
        "@id": "http://codh.rois.ac.jp/pmjt/book/200003067/manifest.json",
        "@type": "sc:Manifest",
        "label": "唐糸草紙"
      }
    }
  ]
}
```

- **キュレーション**：自らの関心に沿う資料を収集し、配列して展示する行為。
- **IIIFの現状**：物理的な資料（Manifest）を**断片化**し、論理的に**再構成**できない。
- **IIIFの拡張**：既存規格の自然な拡張となる**Curation API**を考案。



IIIF と Curation

- Presentation APIの外側に**Curation** ノードと **selections**、 **within** リンクを付加。
- Collection以下は提供者側が定める固定的な構造 **Curationは外部から誰でも自由に編集可能な構造。**

<http://iiif.io/api/presentation/2.1/>

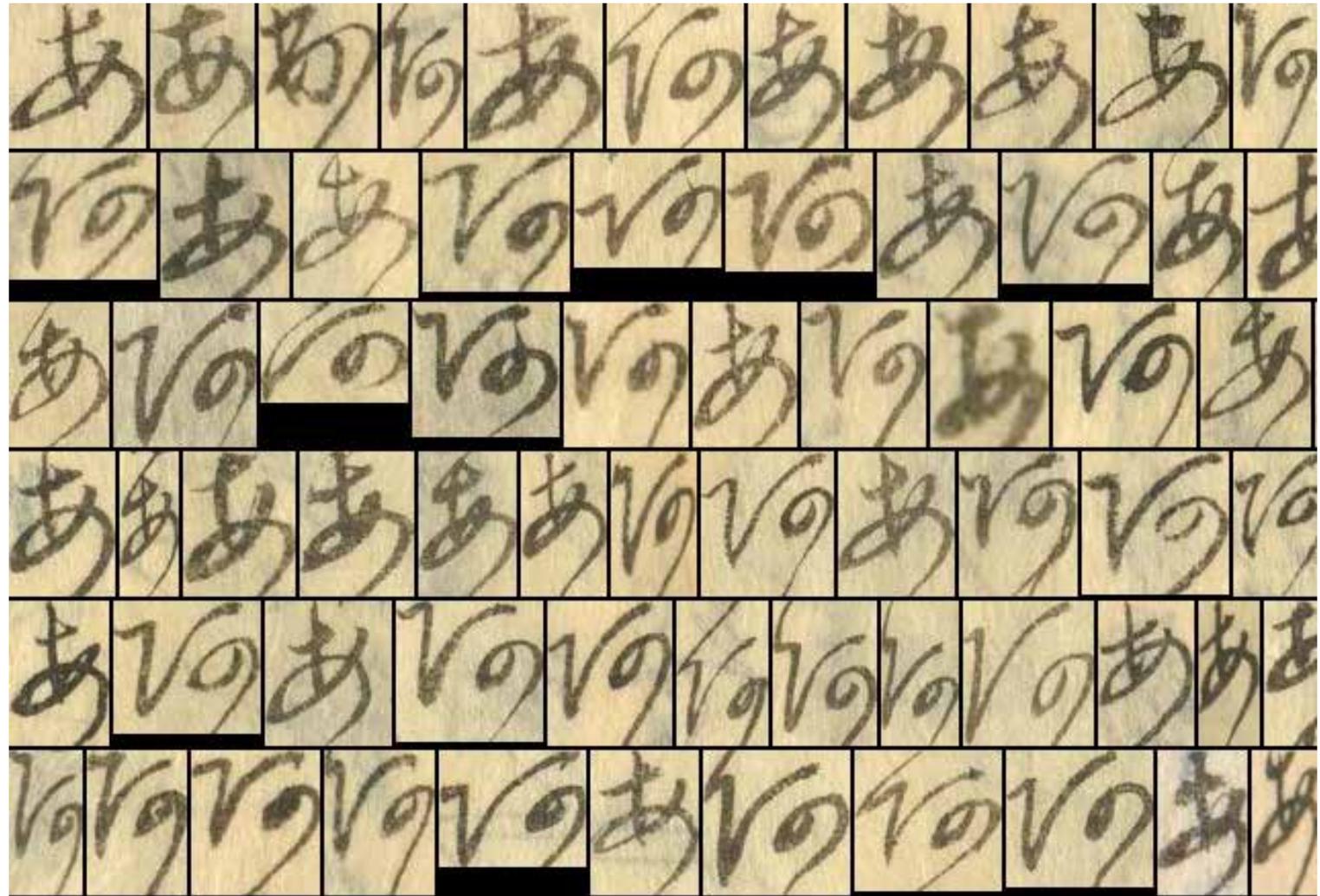
識別子の付与

- 学術情報は、**グローバルな識別子で結ばれる知識の網**になりつつある。
- DOI (Digital Object Identifier) : 多くの学術データの基本的な識別子。
- **国文研** : 古典籍の「書籍」に対するDOIを、国文研書誌IDをベースに付与。
- **CODH** : オープンデータの構造をDOIと揃え、派生データに対してDOIを付与する。

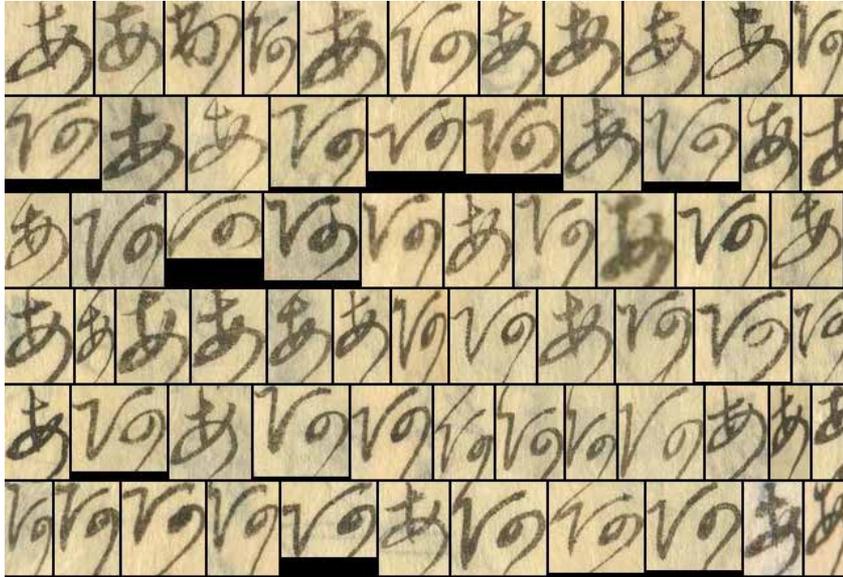
2. 日本古典籍字形データセット

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

日本古典籍字形データセット



人間のための学習データ



- 字形のバリエーションを目で見て確認できる。
- 学習アプリの素材データにもなる。
- くずし字を読める人を増やすことが、データ利活用の好循環を生み出す。

機械のための学習データ

文字種	文字数
し	3,929
に	3,147
の	2,908
て	2,398
り	2,193
を	2,021
か	1,910
く	1,739
き	1,715
も	1,463
1,521文字種	86,176文字

- 機械学習用のデータセットとして利用。
- **深層学習**ライブラリ Kerasを用いたCNNサンプルプログラムを配布。
- **座標情報**が存在するため、文字単位を越えた問題設定も可。
- **字母**の扱いは将来課題。

スク립トーム解析へ

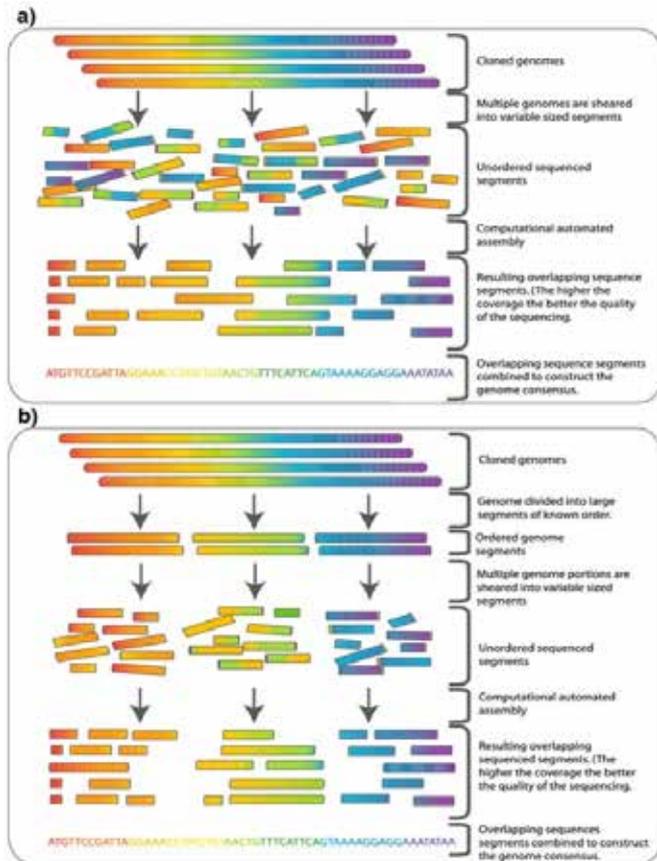
- **くずし字OCR**：文字の切り出しを手で行う必要がある。
- **画像検索**：文字の切り出しは必要ないが、どこかでテキストに変換したい。
- **鶏と卵の関係**：分割と認識が相互に依存する状況を乗り越えて全自動化したい。
- **スク립トーム解析**：古典籍の全体を解読するため、あらゆる手段を総動員する。

ゲノム解析の歴史

我々の現在地？

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）？

断片を読んでつなげる



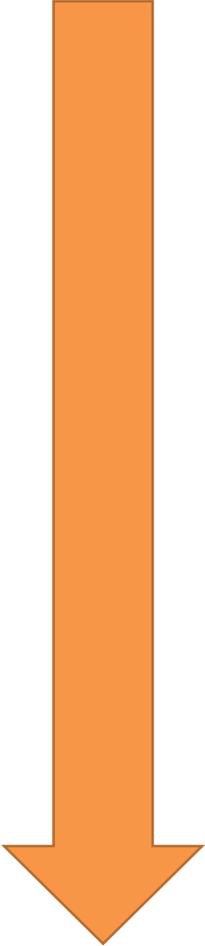
- 全体をいったん細かい断片に切り刻む。
- 断片の遺伝子配列（文字列）を解読する。
- 文字列の重なりを活用して断片をつなげる。
- このアルゴリズム開発が、ゲノム解読の鍵。

Commins, J., Toft, C., Fares, M. A. - "Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects." Biol. Procedures Online (2009). Accessed via SpringerImages, [CC BY-SA 2.5](#)

3. 江戸料理レシピデータセット

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

江戸料理レシピデータセット

- 
1. 江戸の料理本をデジタル化
 2. くずし字を翻刻
 3. 翻刻を現代語訳
 4. 現代語訳をレシピ化・公開
 5. クックパッドでもレシピ公開
 6. つくれぽで個人の経験を共有

協力：合同会社AMANE

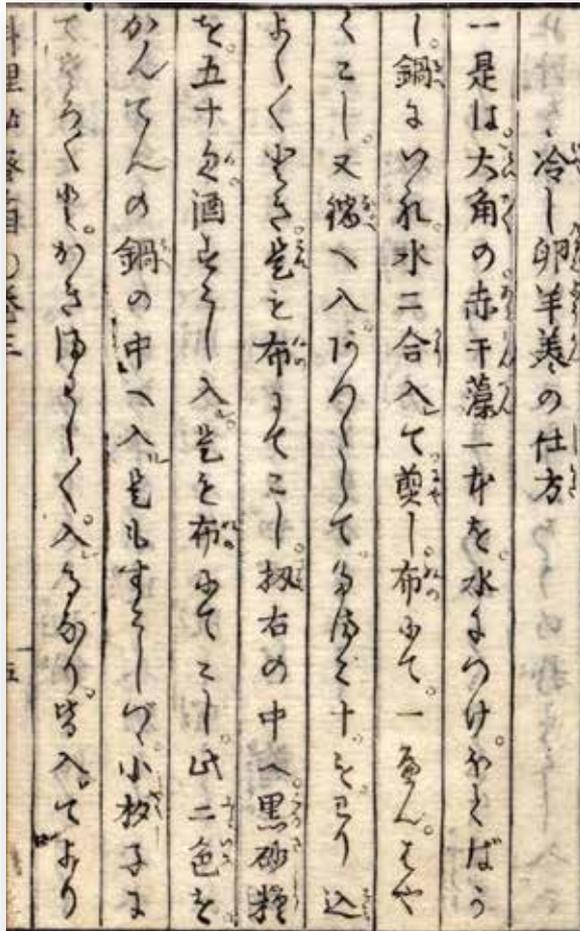
1. 江戸の料理本をデジタル化



- 『**万宝料理秘密箱**（まんぼうりょうりひみつばこ）』
- 初版は1785（天明5）年。
- 素材としての卵一種で100の料理のバリエーションを楽しむ「**卵百珍**」。
- 「**百珍物**」：食べるための食事から、食を楽しむ生活への変化という時代背景。

日本古典籍データセット
（国文研所蔵）

2. くずし字を翻刻

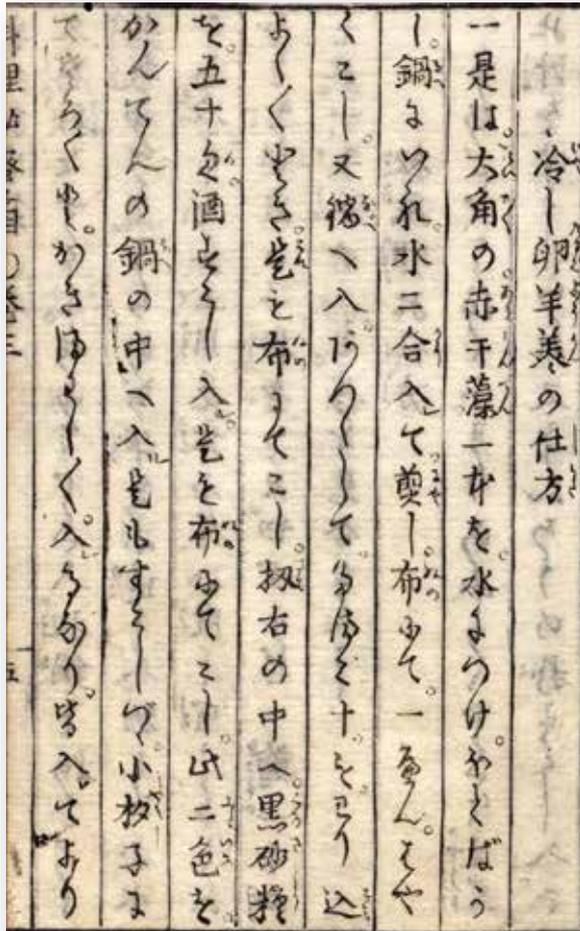


日本古典籍データセット
(国文研所蔵)

1	是は大角の赤干藻一本を水につけほとばかし
2	鍋にいれ水二合入して煎し布にて一へんはやくこし又鍋へ入れあつくして
3	たまご十ウをわり込よくよくとき是も布にてこし
4	扱右の中へ黒砂糖を五十匁酒すこし入ル是も布にてこし
5	此二色を かんてんの鍋の中へ入ル
6	是もすこしづつ小杓子にてそろそろとかきまわしかきまわし入れるなり
7	皆入してより又葛粉をすこし水にてとき入レ
8	扱鍋をぬき早く折敷にてもうちあげ平めに延し入レ物ともに水に入レ冷し遣ふ

江戸料理レシピデータセット (CODH制作)

3. 翻刻を現代語訳



日本古典籍データセット
(国文研所蔵)

2016/12/10

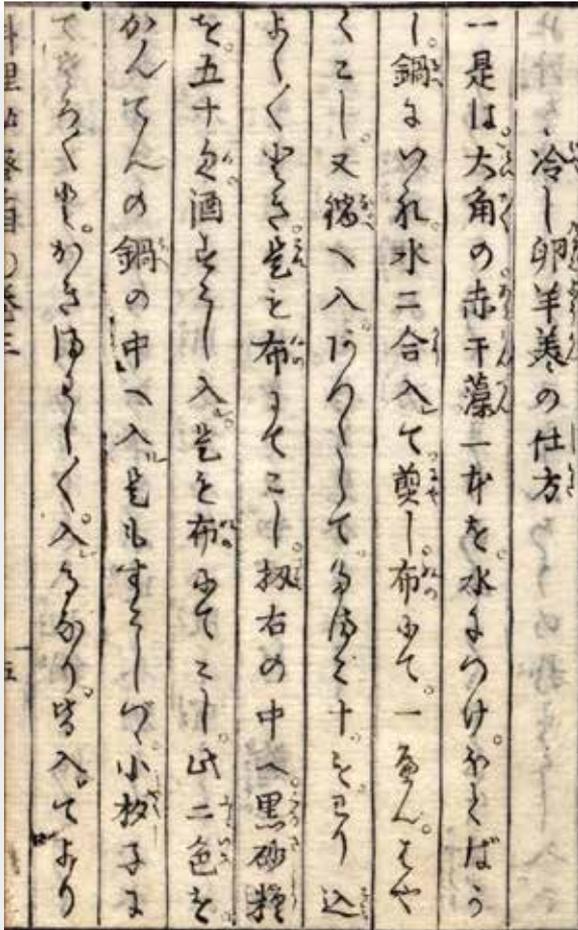
1	大きな赤寒天を1本水に付けてふやかす。
2	鍋に寒天と水2合(360cc)を入れて煮溶かす。
3	を一度布で素早く漉し、再び鍋に入れて熱する。
4	生卵10個をよく溶き、布で漉す。
5	の中に黒砂糖50匁(200g)と酒少しを入れ、布で漉す。
6	を寒天の鍋に入れる。小さな杓子で少しずつそろそろと混ぜながら入れる。
7	を全て鍋の中に入れたら、葛粉を水で溶き、鍋に入れる。
8	鍋を火から上げ、素早く中身を容器(折敷)に広げ、平たく延ばし、容器ともに水で冷やす。

江戸料理レシピデータセット(CODH制作)

じんもんこん2016

25

4. 現代語訳をレシピ化



日本古典籍データセット
(国文研所蔵)

1	寒天を水につけて、ふやかします。
2	生卵をよく溶きます。
3	溶いた生卵を布でこします。
4	黒砂糖と酒を入れ、溶かします。
5	4を3に入れ、再びこします。
6	鍋に寒天と水(180cc)を入れて煮とかします。
7	6を布などでこし、再び鍋に入れて熱します。
8	7の熱した寒天の中に、5の卵液を少しずつ入れます。
9	全て入れ終わったら、水でといた片栗粉を鍋に入れてさっと混ぜ合わせます。
10	鍋を火からあげ、中身を容器に入れます。
11	冷蔵庫で、2時間程度冷やします。

江戸料理レシピデータセット (CODH制作)

文字と写真による調理手順



江戸料理レシピデータセット (CODH制作)

現代の状況に合わせたレシピ

- **材料の違い**：「葛粉」を「片栗粉」へと、入手しやすい材料に変える。
- **道具の違い**：「容器ごと水で冷やす」を「冷蔵庫で2時間程度冷やす」へと、現代の道具に合わせて変える。
- **分量の違い**：「卵10個」を「卵5個」へと現代の生活にあった分量に変える。

現代の文化を反映したレシピに「翻訳」

過去のことば・道具・文化

- **ト活卵**：蒲鉾を藁の太さ程に切る。
- **小豆餅卵**：ゆで卵を水飴の中に入れ、表面にくるりと塗る。
- **鶉卵**：鶉か鳩か　か雲雀か鴨か、その他どのような鳥にしても、鳥肉を丁寧によく叩く。
- **家主貞良卵**：上には、行灯の火皿を乗せ、この中に灰をうすく引き、火を入れて焼く。

江戸時代の文化を感じさせる記述から、過去の生活への想像を膨らませる。

江戸料理レシピデータセット

日本古典籍データセットに含まれる江戸の料理本を、現代の生活にも取り入れるために、現代レシピに変換して提供します。

最初の江戸料理レシピとして、100種類以上の秘料理を集めた『万宝料理秘伝集 御百珍』を取り上げます。

「万宝料理秘伝集 御百珍」の江戸料理レシピ

くずし字を読める日本人が少ないという中で、日本古典籍データセットのようなデジタル画像を提供するだけでは、市民によるオープンデータ活用を進めることは難しいのが実情です。古典籍を日常生活にどのように活用していけばいいか、と考えているところで思い当たったのが江戸時代の料理本でした。これを現代語訳すれば現代でも料理を作って楽しめるのではないかと考えました。

雑煮などの季節の料理や地方色豊かな料理などは、日本人の生活に深く根ざしたものです。そして日本の料理としての和食は、単なる料理法を超えて自然の尊重という日本人の精神に基づく文化を表すとも言われています。平成25年には「和食：日本人の伝統的な食文化」がユネスコ無形文化遺産に登録され、和食文化に対する国際的な認知度も高まってきました。そんな和食という自身の文化をより深く理解するには、過去の料理について学び、気が向けば作ってみることのできるようなレシピデータが必要だと考えました。そこで以下のような「レシピ化」のプロセスに取り組みました。

データ概要

原本画像データ	日本古典籍データセットで公開する画像です。くずし字を読み、かつ江戸時代の日本語や料理法を知っていれば料理が作れます。
翻訳テキストデータ	原本画像のくずし字をテキスト化したデータです。江戸時代の日本語や料理法を知っていれば料理が作れます。
現代語訳データ	翻訳テキストデータの内容を現代の日本語に翻訳したデータです。江戸時代の料理法を知っていれば料理が作れます。
現代レシピデータ	現代語訳データの内容を、現代の道具や食材でも作れるものに実装し、食材の分量や写真を加えてより具体化したデータです。手順に従えば料理が作れます。

1. 翻刻：全107点
2. 現代語訳：翻刻107点中20点
3. レシピ化：現代語訳20点中5点

オープンデータ (CC BY-SA)
としてウェブサイトで公開。

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

クックパッドでもレシピ公開

クックパッドと日本家政学会 食文化研究部会が運営する「クックパッド江戸ご飯」に参加。



多くの人々が既に馴染んでいるウェブサービス（アプリ）からも、江戸料理のレシピを活用。

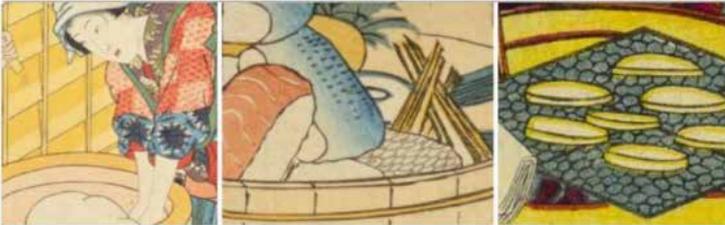


<http://cookpad.com/recipe/4153357>

極めて大きな反応

人文学オープンデータ共同利用センターさんがリツイート

うずら @caille2006 · 11月26日
このプロジェクトがすごいのは、古文書の情報をさらに現代の生きた情報にするために、クックパッドにアカウントを開設してレシピを公開し「つくれば」も受け付けていること。江戸ご飯とつくればというこの未来感パネい。 cookpad.com/kitchen/146046...



クックパッド江戸ご飯 のキッチン

プロフィール

トップ	レシピ 32	つくれば 0	献立 0
-----	-----------	-----------	---------

レシピを検索

7478 リツイート

<https://twitter.com/caille2006/status/802575840819089409>

2016/12/10

人文学オープンデータ共同利用センターさんがリツイート

NII 国立情報学研究所(NII) @jouhouken · 11月24日
[プレスリリース]
江戸の文化を現代に取り込む「江戸料理レシピデータセット」を整備～江戸時代の料理本を「レシピ化」し、クックパッドでも公開～
nii.ac.jp/news/2016/1124



← 1 1,074 971

1074 リツイート

<https://twitter.com/jouhouken/status/801693251052781568>

じんもんこん2016

32

作って食べてみた

0テレ NEWS24

2016年(平成28年) 12月6日(火) 今日の実況 仙台

UR賃貸住宅 社宅も UR

礼金 ¥0 手数料 ¥0 更新料 ¥0 保証人 不要

トップ 社会 エンタメ 国際 政治 経済 スポーツ 投稿 ブログ 特集 地方 天気 地震 防災

トップ > 社会 > 「江戸ご飯レシピ」再現、作って食べてみた

0テレ NEWS24 「江戸ご飯レシピ」再現、作って食べてみた

ツイートする シェアする 2016年11月30日 21:05

江戶ご飯レシピ Hot Word

全文

11月28日、「江戸ご飯レシピ」という言葉を含んだツイートが大きく伸びました。

「いま見るべき」「いま知るべき」厳選ニュースをメルマガでお届け

ニュース検索 検索

カレンダー検索

★ アクセスランキング 社会

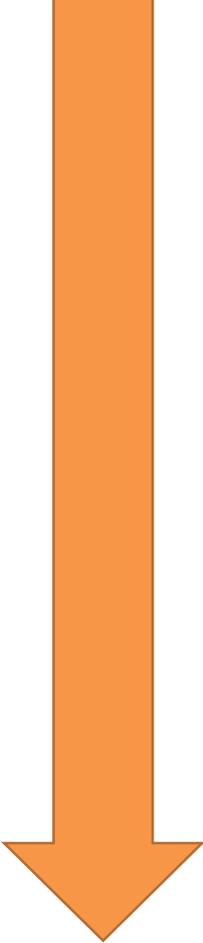
- 1 豊中市妊婦刺殺「殺意はなく事故」と否認
- 2 電車で女性の髪切り「アクションで売る」
- 3 千葉大医学部3人の英名公表 医師も逮捕
- 4 女性の祥英橋「ばらまくぞ」航空自衛官逮捕
- 5 東京の異端スタッフ祭り「おっぱい祭り」開催

<http://www.news24.jp/articles/2016/11/30/07347892.html>

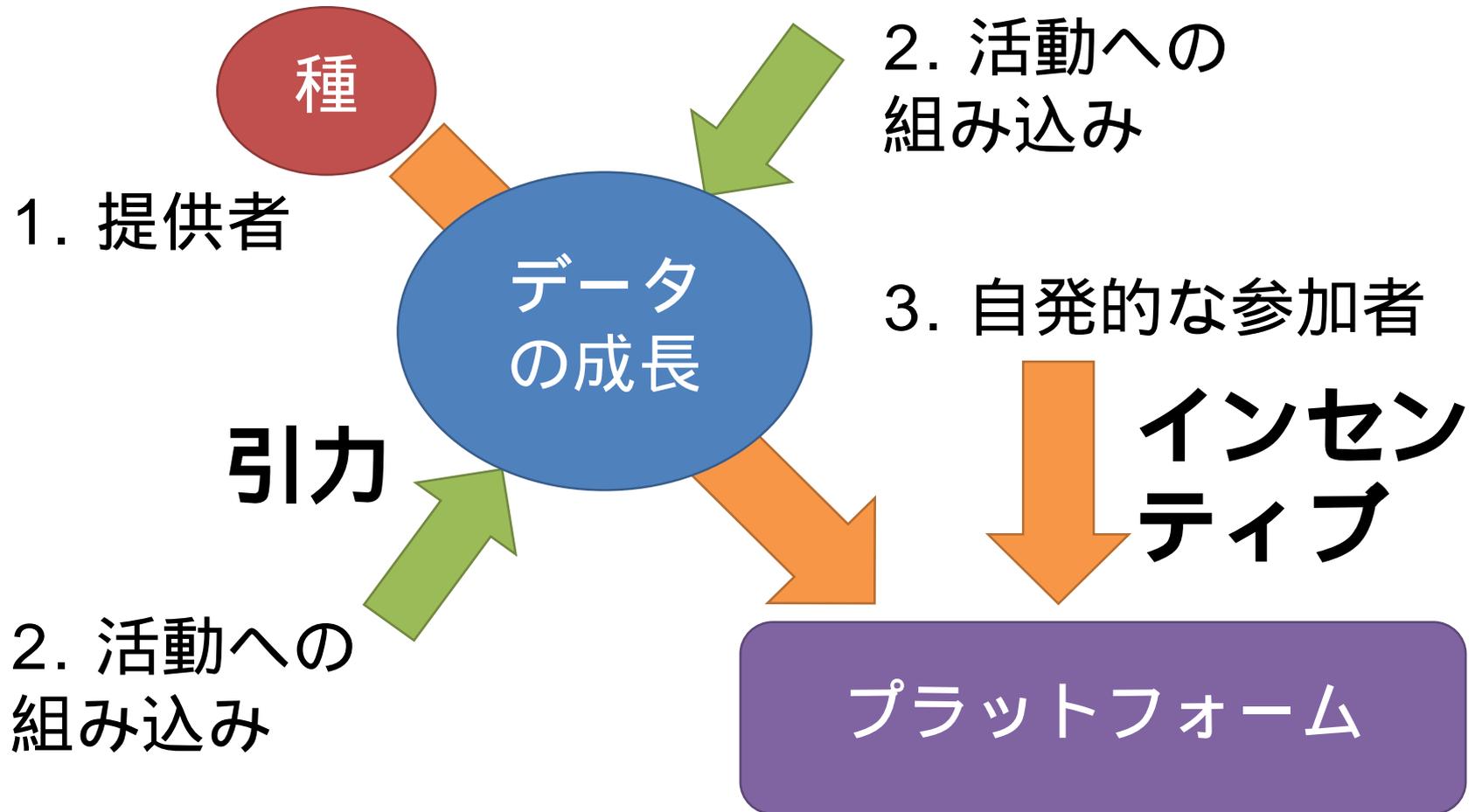
得られた教訓

- 「データをクックパッドでも公開する」と説明すると、**人々の顔がぱっと輝く**。
- 仮説：クックパッドという現代と江戸料理という**過去のギャップ**に**意外感**あり。
- 仮説：**馴染み**のプラットフォームに入ると、「**自分たちのデータ**」と感じる。
- **市民のためのオープンデータ**：提供者側から積極的に距離を詰めていくべき。

アイデアの実現化

- 
- 2015年12月 アイデアソン「じんもんそん」を開催。ここで料理本の存在を初めて知り、今回のアイデアを発案する。
 - 2016年1月 クックパッドを訪問し、協力を依頼。
 - 2016年11月24日 江戸料理レシピデータセット + クックパッド公開。
 - **2016年12月 アイデアソンを開催！**

オープンデータ化プロセス



おわりに

データ駆動型
サイエンス

深める

研究者

増やす

機械

オープン
サイエンス

市民参加型サイエンス
(シチズンサイエンス)

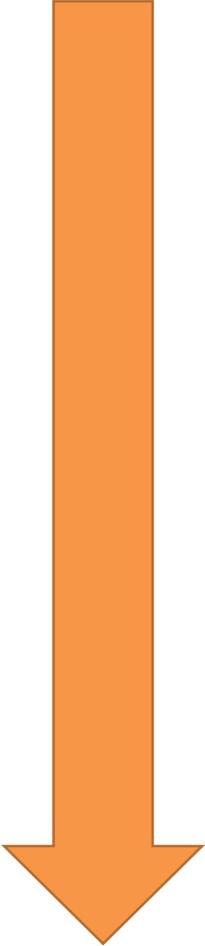
人間と機械の競争
vs. 人間と機械の協
調(チームング)

市民

超学際的データ
プラットフォーム

広げる

データからコミュニティへ

- 
1. オープンデータを公開する。
 2. コンテストで成果を競う。
 3. コミュニティで知識を共有する。
 4. データを扱える人材がどんどん育っていく。

まとめ

- CODHと国文研の協働による、日本古典籍関連のオープンデータを紹介した。
- オープンサイエンスへの多様な参加者（研究者・機械・市民）を巻き込んだ。
- 「江戸レシピデータセット」の反応が大きかった理由は、さらに分析したい。
- 日本文化の網羅的解析という野心的目標に向けて、できる部分から取り組む。