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・システム研究機構人文学オープンデータ共同利用センター）
山本 和明（国文学研究資料館）

人文学におけるオープンサイエンスを推進するためには，データをオープン化することでどのような新しい可能性が生まれるのかについて，目に見える形で示すことが重要である．そこで本論文は，我々が進める「歴史的典籍 NW 事業」のオープンデータから派生する 3 種類のデータセットを紹介する．これらは人文学者向き，情報学者向き，市民向きと利用者が大きく異なるため，ニーズと期待される成果に応じてデータセットの構築方法をどう工夫すべきかを論じる．さらにオープンデータの一環となる画像公開では，キュレーションを可能とするように既存の IIIF 規格を拡張する Curation API 規格を提案し，ユーザが興味に応じて自由に本を断片化して再構成できる環境の実現を目指す．

Construction of trans-disciplinary data platform that explores open data in the humanities

Asanobu KITAMOTO (National Institute of Informatics / Center for Open Data in the Humanities, Research Organization of Information and Systems)
Kazuaki Yamamoto (National Institute of Japanese Literature)

To promote open science in the humanities, it is important to visualize what kind of new possibilities we can explore from the openness. Hence this paper introduces three types of datasets derived from open data created by the “network project for pre-modern Japanese text.” Those datasets have different users, such as humanists, computer scientists, so the paper discusses how to design the datasets to reflect user’s needs and expected results. Moreover, in terms of the image browsing platform associated with the open data, we propose Curation API as an extension to the existing IIIF specifications to enable users to enjoy curation, which amounts to the fragmentation and reconstruction of books based on the user’s own interest.

１．はじめに

オープンサイエンスという概念が広まるにつれ，多くの学問分野においてデータのオープン化が進展し，人文学にもその動きが徐々に浸透しつつある．例えば国文学研究資料館が中心となって進める「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」（歴史的典籍 NW 事業）[1] は，歴史的典籍に関する大規模データのオープン化とデータを対象とした国際的なコラボレーションを目的としており，オープンサイエンスの考え方と親和的な目標を掲げている．

歴史的典籍をネット上で誰でも閲覧できる環境は，人間関係等の条件や地理的な距離とは無関係な利用を可能とし，アクセスコントロールが権力化することを防ぐという意義がある．また，同じ時間に多くの研究者が同じデータにアクセスできることは，研究への参加条件を公平にし，コミュニティの研究レベルを高める効果が期待できる．さらに研究者だけでなく市民もデータ活動に積極的に参加する道を開くことで，オープンデータがより大規模化かつ多様化する可能性も生まれる．そして，第三者によるデータの検証を可能とすることは，研究の透明性向上にも寄与する．

このようにオープンサイエンスには，（再）利用，透明性，参加という 3 つの軸が存在する．そこで本論文は，オープンサイエンスの考え方を反映した超学際的データプラットフォーム構築について，情報・システム研究機構 人文学オープンデータ共同利用センターと国文学研究資料館が協力して取り組む，歴史的典籍 NW 事業を元にしたオープンデータの構築を中心に論じる．

このプラットフォームでは，1）日本古典籍データセット，2）字形データセット，3）江戸料理レシピデータセット，という 3 種類のデータセットを最初のオープンデータとして公開する．しかしこれらは，1）主に人文学者向け，2）主に情報学者向け，3）主に市民向け，と想定する利用者が大きく異なるため，それぞれのニーズと期待される効果に応じてデータセットを設計しなければならない．そこで本論文は，オープンデータ構築の過程で生じる具体的な問題の解決法を論じつつ，オープンデータ設計の背景となる考え方や将来への展望なども俯瞰的に紹介する．

2．日本古典籍データセットの公開

2.1　日本古典籍データセットとは

歴史的典籍 NW 事業でデジタル化された画像は，これまで国立情報学研究所の情報学研究データリポジトリ（IDR）にて，「国文研古典籍データセット」として 350 点を先行公開していた．2016年11月にこれを「日本古典籍データセット」と名称変更して人文学オープンデータ共同利用センター（CODH）に移し，点数も 700 点（158,159 ページ）に増やして正式公開した．データのライセンスには引き続きクリエイティブ・コモンズのライセンスである CC BY-SA 4.0 を採用し，二次利用を歓迎するオープンデータとして提供する点は従来通りである．

データ提供の単位は，書籍としてのまとまりを一括りとする．そして，デジタル化した画像に加え，書誌に関するメタデータや専門家が付与したタグデータなどの CSV ファイルを用意し，それらを一つにまとめて ZIP ファイルに固めて提供する．ただし IDR サイトでは ZIP ファイルの中身を確認する機能がなかったため，ダウンロード前にデータを選ぶことができなかった．そこで CODH サイトでは，画像の中身を確認してからダウンロードするための画像ビューアーを設置することとした．そのために設計，実装したのが IIIF Curation Viewer である．

2.2　Curation API の提案と Viewer の実装

IIIF（International Image Interoperability Framework）とは，画像へのアクセスを標準化し相互運用性を確保するための標準を定める国際的なコミュニティ活動である．これまでの成果物は，画像へのアクセスを定める IIIF Image API 2.1，書籍などの構造を定める IIIF Presentation API 2.1，そして検索に基づくアクセスを定める IIIF Search API 1.0 という 3 つの API である．API は仕様が公開されているため，誰でも自由に API に準拠したソフトウェアを開発し，オープンソースとして公開できる．そのため IIIF に対応したオープンソースソフトウェアが既にいくつも誕生しており，これが IIIF の使い勝手を向上させることで，さらにユーザが集まるという好循環が働いている．

ただし IIIF はまだ未成熟の規格であり，重要なユースケースをすべてカバーできているわけではない．例えば IIIF Presentation API 2.1 仕様の冒頭には，この仕様が物理的なオブジェクトの画像表示を対象とし，ページのナビゲーションや各種テキストの表示はスコープ内とするが，興味のあるデジタルオブジェクトを発見したり選択したりする機能はスコープ外とすることが明記されている．このように IIIF がカバーできていないユースケースに対しては，現状の API の範囲で満足するのではなく，IIIF の世界観を踏まえた新しい規格の提案をすることが，コミュニティへの貢献として期待されていると考える．


```
{
  "@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"
      }
    }
  ]
}
```


図 1　Curation API に従う JSON の例．

特に我々が着目したユースケースは，興味のあるデジタルオブジェクトを複数の書籍から複数選ぶという選択機能である．これはテーマごとに資料を収集し配列するキュレーションには必須の機能であるが，現状の IIIF 規格の枠内ではこうした機能を実現することは困難である．なぜなら IIIF Presentation API は物理的な書籍（オブジェクト）を基本単位とするため，それをバラバラに断片化し別の視点で再構成するというユースケースは含まないためである．そこで我々は，キュレーションを実現するための新しい規格として，　Curation API を提案することとした．

キュレーションという言葉は，もともとのアートの世界においては，キュレーターの独自の視点に沿って作品を選択し配列する行為や，その価値

を外に向けて論じる行為を意味していた．しかし近年はキュレーションの意味も大きく広がりつつあり，そのための Curation API にも様々なユースケースが想定できる．しかし本論文は，最も基本的な機能として，複数の書籍からリソースを収集してリスト化する機能を検討した．

ここで重要となるのが IIIF Presentation API における Manifest タイプの扱い方である．Manifest タイプは物理的なオブジェクトの構造を固定的に表すものである一方，Curation タイプは利用者側の視点によって流動的に変化する．そこで，Manifest という物理的かつ固定的な構造の外側に，Curation という論理的かつ個人的な構造を付け足していく機能が必要となる．そのためには，Manifest の外側から Manifest の内部の要素を指したリストを作成し共有する機能が必要となるが，これは IIIF Presentation API の枠内では実現できない．そこで Curation API という新たな規格を考案することとした．

ここで課題となるのは，IIIF Presentaiton API で定義された Manifest の中の任意の Canvas を，その外側からリスト化して Manifest を越えたセレクションを実現する記法である．そこで，最上位に来る Curation タイプは，hasRanges プロパティを通して Range をリスト化し，hasRanges プロパティを JSON-LD の context.json を通して selections と表記することで，モデル上での意味を明確にした．そして，Range タイプが isPartOf プロパティを持てるようにモデルを拡張することで，各 Range が属する Manifest を指せるようにした．これによって，複数の Manifest から選択した Canvas をまとめることができた．

なお，上記の isPartOf を JSON-LD で within と表記するのは，他のタイプに対する利用と同じ意味である．IIIF Presentation API には名詞形プロパティと前置詞系プロパティが存在するが，前置詞系プロパティは小さなクラスから大きなクラスを指す場合にのみ使われているように見える．そこでこうした使い分けの慣習に従って Curation API のプロパティを定義した．

このように Curation API は，既存の IIIF 規格の自然な拡張であるため，他の IIIF 規格と組み合わせて利用できる．そこで我々は，Curation API の参照実装となる IIIF Curation Viewer を開発し，オープンソースライセンスで公開した．このビューアーは，(1) Curation API に従う JSON ファイル（図 1）を URL にパラメータ指定する，(2) 選択するページなどの情報を URL パラメータに指定する，という 2 つの方法でキュレーションの対象を指定できる．(1)は複数の提供元の資料をキュレーションでき，関連情報などを持たせることもできる．一方(2)は単一の提供元の資料のキュレーションに限定されるが，JSON ファイルを別途用意する必要がなく，選択ページの情報を簡易的に指定できる利点がある．

2.3 識別子の付与

データ公開におけるもう一つの重要な作業は，永続的な識別子の付与を通した安定的なアクセスの確保である．IDR の試行公開版では専用の識別子を利用していたが，CODH の正式公開版では国文研の国書総目録に紐付けられた識別子である「国文研書誌 ID」を，安定的な識別子として利用することにした．この識別子を様々なオープンデータに活用すれば，今後のデータ統合が簡単になる効果が期待できる．さらにこの識別子は，国文研が付与する DOI（Digital Object Identifier）にも利用する予定であり，このグローバルな識別子によって世界の学術情報プラットフォームとの相互運用が実現する．国文研は JaLC（ジャパンリンクセンター）の正会員となり，CODH は国立情報学研究所が運用する JAIRO Cloud の実験プロジェクトに参加するという方法で，DOI を付与する準備を進めている．今後は様々なデータについて，DOI を活用する計画である．

3．字形データセットの公開

3.1 字形データセットとは

字形データセットとは，歴史的典籍をデジタル化した画像の中に含まれる文字を，一文字ずつ切り出して作成したデータセットであり，1）原本補正画像，2）文字座標データ，3）字形画像，4）作業文書の 4 種類から構成される．文字数は 2016 年 11 月現在では約 8 万文字であるが，2016 年度中には 40 万文字に増やすことを目標に，作業を進めている．その主な利用目的は，情報学者による機械学習のための学習データセットである．これを活用して文字認識ソフトウェアを開発できれば，最終的には人間が翻刻するよりも早い速度で大まかな翻刻テキストが作れるかもしれない．

文字座標データは，原本補正画像の中に含まれ，かつ読める文字を，切り出してリスト化したものである．基本的な構造は（原本補正画像 ID，文字座標（xywh），Unicode コードポイント，ブロック ID，文字 ID）の 5 つ組となっており，文字座標を使って原本補正画像上のすべての文字の位置を把握できる．もし字形だけに関心があるなら，原本補正画像から切り抜いた字形画像だけを配布すればよい．しかしこれを機械学習のための学習データセットとして使う場合，前後文字の文脈を把握したり，文字よりも大きな単位で切り出したりするニーズが予想されるため，文字座標データも含めてデータセット化することとした．逆

図 2　字形データセットの例．同じ文字に対する字形のバリエーションを一覧できる．

に翻刻テキストは，機械学習にとって必須ではないとの考えに基づき，データセットには含めない．

次に原本補正画像は，2 章でオープン化した古典籍画像を対象に前処理を施したものである．具体的には，見開き画像を分離し，画像を回転して正立させるという前処理を適用し，作業の負荷を低下させるとともに，学習データセットとしての品質を向上させる．また字形画像は，2)の文字座標データの文字座標(xywh)で原本補正画像を切り抜いたものである．厳密に考えれば 1)と 2)があれば 3)の字形画像は冗長であるが，字形だけに関心がある利用者の利便性を高めるために用意した．最後に作業文書は，読めない文字などに関する情報をまとめる．字形データセット作成では，国文学研究者ではないがくずし字の読解に熟練した人物が作業を担当するため，その精度は十分に高いと評価できる．しかしどうしても読めない文字が残るため，作業文書にその記録を残すことで，将来のデータ修正のための参考情報とする．

なお Unicode コードポイントについては，翻刻テキストをベースとして変体仮名の Unicode は用いない．ゆえに字形データセットからは変体仮名の字母情報が欠落しているが，人手または機械による字母情報の付与は今後の課題とする．

これまでも国文学研究では多くの翻刻テキストが作られてきた．紙の書籍の翻刻では一文字ずつ書いていくしかないが，デジタル画像の場合は専用または汎用の作業ツールを使い，ツール上で文字を四角で囲んでコードを与えるという作業を繰り返すことが多い．このとき，翻刻という最終成果物のみに関心を持つ人文学者は，座標情報という中間生成物を丸ごと捨てていたが，実は機械学習研究者にとっては，これこそが欲しいデータなのである．利用者の関心によってはゴミが宝に変わりうるため，データセットの構築にあたってはニーズをよく把握することが重要である．

3.2　文字認識からスクリプトーム解析へ

本データセットの波及効果として期待するのは，光学的文字認識（OCR）ソフトウェア開発である．具体的な課題は，1) 画像からの文字切り出し，2) 文字コードの推定，3) 文脈情報による改善，の 3 点である．まず字形データセットの貢献が期待できるのは 2)の部分である．ある文字コードに属する多くの字形画像をデータから学習できれば，その文字コードと別の文字コートとを識別する学習機械の精度が向上することになる．特に様々な分野で革命的な精度向上を示す深層学習（ディープラーニング）では，データが多ければ多いほど精度が向上する傾向があり，これによって新しい OCR が実現できる可能性がある．

しかし注意すべきことは，文字認識の完全な自動化を達成するには，課題 1)も重要という点であ

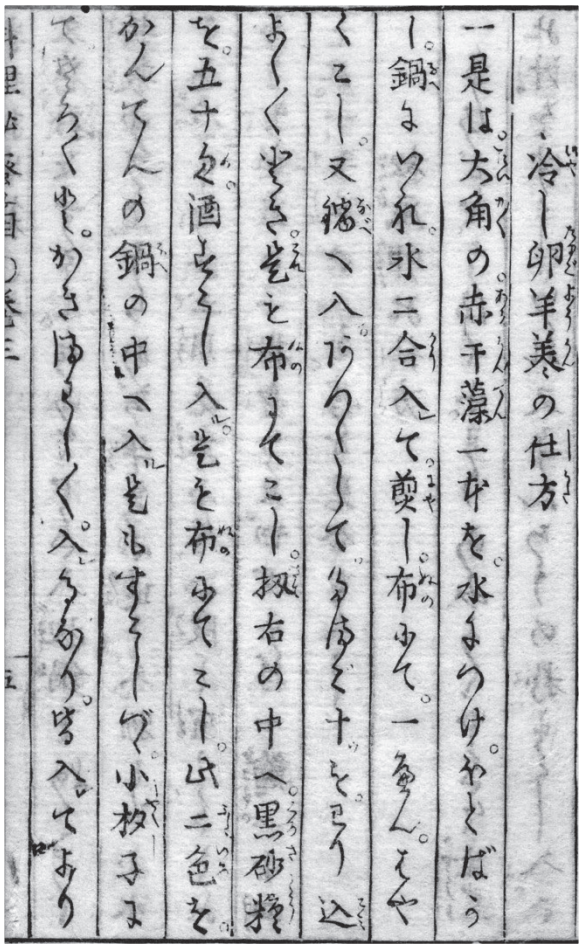

図3 『万宝料理秘密箱』(国文研蔵)「冷し卵羊羹」をくずし字で記述した原本画像.

る．現代の印刷文字では，文字間の空白や行・列の認識を活用して文字を切り出せるが，くずし字は文字間の空白がなく連続しており，行・列も鉛直水平ではないため，文字の切り出しはより難しくなる．このようなセグメンテーション（分割）は，画像処理における本質的な問題である．

課題 1)と課題 2)の関係は，いわば鶏と卵の関係にある．認識できれば分割すべき場所もわかるが，認識にはまず分割が必要だからである．人間が文字の切り出しを担当し，機械が文字を認識するという現行の方法は，人間と機械が得意な作業を分担するチームを組むことで，鶏と卵の関係をある程度は回避している．しかし歴史的典籍 NW 事業のように，30 万点の大規模画像を網羅的に解析するためには，やはり全自動的な処理を長期的な目標に据えるべきではないかと考える．

そこでこの問題に挑むための考え方として，スクリプトーム解析（scriptome analysis）[2]という考え方を提唱したい．これは過去のゲノム解読などの歴史に学び，記号列の全自動かつ網羅的な解読に向けた長期的な方針に関するコンセプトである．文字認識ではこれまで，先頭から 1 文字ずつ読んでいく方法が試みられてきた．一方ゲノム解読では，ゲノムという長い記号列を細かい単位に断片化し，断片を読んで後からくっつけるショットガン法のような新しいアイデアを組み合わせて，全ゲノム解読に成功した．また古代文字のヒエログリフ解読では，ファラオの名前が記されたカルトゥーシュという特別な部分の解読にまず成功し，そこから徐々に解読範囲を広げていくことで，全ヒエログリフ解読に成功した．こうした過去の成功例を踏まえると，読めるところから読み，欠けた部分は他のアルゴリズムも総動員して補っていくような，これまでとは異なるアプローチにも検討の価値がある．

ヒトゲノム解読は当初は 100 年かかるプロジェクトとも言われていたが，機械的な解読方法が提案され，それが爆発的に発展したことから，本格開始から 15 年程度でヒトゲノムの全体を解読できた[3]．また，最初のヒトゲノム解読には全世界で約 3000 億円の費用を要したが，現在ではそれが数十万円程度に低下するなど，プロジェクト開始当初には想像もつかなかった技術革新が生じている．古典籍についても全解読は現在のところは夢物語だが，深層学習等の機械学習技術の展開によっては，10 年単位で見れば画期的に進展する可能性もある．そうした知見を研究コミュニティで継続的に共有するため，評価型ワークショップの企画・開催にも取り組む構想がある．

3.3 機械の学習と人間の学習

字形データセットは主に機械が学習するためのデータセットとして構築しているが，これは人間が学習するためのデータセットとしても使える．このデータセットを使えば字形のバリエーションを簡単に一覧できるため，学習者はより多くのくずし字に触れて字形の実際の変化を把握できるからである．具体的な方法としては，くずし字学習支援アプリ KuLA [4]などに，字形データセットを取り込む方法を考えている．

字形データセットで字形の概念に触れることで学習が進み，くずし字を読める人が増え，それが将来の字形データセットの充実につながる，というのが字形データセットを核とした好循環の理想形である．日本語の文字でありながら，くずし字を読める人は全国に数千人程度しかいないとも言われる[5]．つまり古典籍を読める人を増やす教育は，潜在的な利用者の絶対数が少ないために翻刻も利活用も進まないという，根本的な問題の解決に寄与する重要な取り組みとなる．機械を賢くするだけにとどまらず，人間を賢くするためのデータセットとなれば，その文化的意義もより大きなものになる．

表 1　『万宝料理秘密箱』「冷し卵羊羹」の翻刻テキスト.

冷し卵羊羹
一　是は　大角の　赤干藻一本を　水につけ　ほとばかし
鍋にいれ　水二合入レて　煎し
布にて　一へん　はやくこし　又鍋へ入レあつくして
たまご十ウを　わり込よくよくとき　是も布にてこし
扨右の中へ　黒砂糖を　五十匁　酒すこし入ル　是も布にてこし
此二色を　かんてんの鍋の中へ入ル　是もすこしづゝ　小杓子にて　そろそろと　かきまわしかきまわし　入レるなり
皆入レてより　又葛粉をすこし　水にてとき入レ
扨鍋をぬき　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

4．江戸料理レシピデータセットの公開

4.1　江戸料理レシピデータセットとは

くずし字を読める市民が少ないという状況で，市民による歴史的典籍の利活用を推進するのは難題である．歴史的典籍を人文学研究の対象として見るなら，過去のテキストを読めれば目的は達成できる．しかし歴史的典籍を市民の日常生活におけるデータ利活用の対象として見るなら，デジタル画像だけでは中身が読めず，美しい絵を楽しむ程度しか活用方法が思いつかない．そこで我々は，歴史的典籍の利活用を推進する鍵は，現在の市民の日常生活と密接に関連するデータセットを提供することにあると考えた．そのためのアイデアが，江戸料理レシピデータセットである．

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あり，和食は単なる料理法を越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言われている．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た．しかし日本人が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もできる情報資源が必要である．そこで古典籍の料理本に出てくる記述を，現代の市民でも使える形式に構造化する作業に取り組んだ．

料理記述の構造化では，以下の点を考慮する必要がある．第一に，料理本の記述と現代のレシピとの違いに関する問題がある．江戸時代の料理本は，お膳の組み合わせとしての献立や，料理名と素材のみの簡潔な記述などが多く，現代のレシピにおいて重要な料理手順に関する記述がほとんどない．第二に，江戸時代と現代の間の言語や語彙，文化の違いに関する問題がある．人文学者が活用するなら翻刻だけで十分だが，市民が活用するには現代語訳が必須である．しかし，料理材料の名前が現代と異なる，当時の道具は入手できない，当時の調味料は現代と味が異なる，といった問題は，現代語訳だけでは解消できない．こうした困難は，料理の難易度を高めてマニアの探究心を刺激する効果はあるものの，すぐ使える情報を欲する一般市民には無用の障害となるため，文化の違いを吸収したレシピ化が不可欠となる．

表 2　『万宝料理秘密箱』「冷し卵羊羹」の現代語訳.

料理名	冷し卵羊羹
食材分類	卵，葛，その他
食材	卵：10 個 赤寒天（大角一本）：8g 葛粉 黒砂糖：約 200g 酒：少し 水：360cc
道具	鍋 布 小杓子 折敷
手順①	大きな赤寒天を 1 本水に付けてふやかす.
手順②	鍋に寒天と水 2 合（360cc）を入れて煮溶かす.
手順③	②を一度布で素早く漉し，再び鍋に入れて熱する.
手順④	生卵 10 個をよく溶き，布で漉す.
手順⑤	④の中に黒砂糖 50 匁（200g）と酒少しを入れ，布で漉す.
手順⑥	⑤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.
手順⑦	⑤を全て鍋の中に入れたら，葛粉を水で溶き，鍋に入れる.
手順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.
使い方	土用見舞には一段と良い．使い方は，刺身物，膾，小皿物，手取肴．その他は，見合わせて使う.

表3 『万宝料理秘密箱』「冷し卵羊羹」のレシピ化．各手順の写真は省略．

タイトル	江戸時代のスイーツ　甘さスッキリ冷卵羊羹
分量（2〜4名分）	卵5個，寒天（赤）1本（4g），黒砂糖100g，水180cc，片栗粉適量，酒適量
1	寒天を水に付けて，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	全て入れ終えたら，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．

そこで我々は，2章で述べた日本古典籍データセットに含まれる『万宝料理秘密箱』の「卵百珍」が述べる卵料理全107点を対象に，レシピ化に向けての構造化を試行した．具体的には，原本画像から開始し，3段階で構造化を進めていく．

1. 原本画像：テキストなし．くずし字を読めないと料理は作れない（図3）．
2. 翻刻：くずし字をテキスト化．江戸時代の日本語がわかれば料理が作れる（表1）．
3. 現代語訳：現代語に翻訳．江戸時代の料理法を想像できれば料理が作れる（表2）．
4. レシピ：食材の分量なども文章と写真で具体的に説明．手順に従えば料理が作れる（表3）．

図表で示す「冷し卵羊羹」の事例を見ても，1や2の情報で料理を作るのは非常に困難だが，4ならば図4の写真で完成形も想像できるため，料理を始める障壁を下げることが可能である．

このように，レシピの構造化では2と4の間に大きなギャップがあり，ここは料理文化に知識を有する専門家の助力がないと埋めることは難しい．こうした専門家の作業も考慮すると，すべての料理をレシピまで構造化するのは難しいことから，段階を進めるにつれて件数を絞ることとした．具体的な件数は，翻刻：107点すべて，現代語訳：20点，レシピ：5点である．

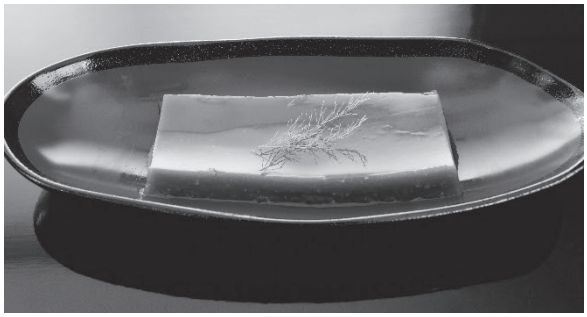

図4 『万宝料理秘密箱』「冷し卵羊羹」の最終完成品を撮影した写真．

4.2 クックパッドのプラットフォーム活用

市民向けのオープンデータでは，データの公開場所も重要となる．CODHウェブサイトで研究データとして公開するだけではなく，市民が日常的に活用するプラットフォームでも公開することが望ましい．そこで着目したのが，クックパッドという日本最大の料理レシピサービスである．ここに江戸料理レシピデータセットを掲載すれば，多くの人の目に触れるのではないか．そこで，クックパッドと日本家政学会 食文化研究部会が進める「クックパッド江戸ご飯」プロジェクトに協力し，ここに現代語訳およびレシピ（写真つき）を投稿することとした．

レシピデータ公開にクックパッドを用いる利点は，料理好きの市民が集まっている点だけでなく，レシピを基に料理を作ったことを報告する「つくれぽ」機能が，江戸料理レシピに関する知識共有のポテンシャルを秘めている点にある．江戸料理レシピから触発された新しいレシピがここで共有されれば，元のレシピにない情報を補えるだけでなく，現代風のアレンジが二次創作的に拡散し，江戸時代の和食の価値を現代の観点から見直すことにつながる可能性もある．

またこの試みは，オープンデータの公開にビジネスプラットフォームを活用するという点でも，我々の今後の展開の試金石となる．ビジネスプラットフォームの利点は，すべてのステークホルダーにメリットがある状況を作り出せれば，ビジネス展開のための堅牢な仕組みの上に乗ることができる点にある．組織の枠を越えてデータという知的財産をオープン化し，超学際的なコラボレーションを活性化させることを通して，オープンデータを核としたオープンイノベーションが起こせるか．学術という世界の外側に出て，異なる目的を持つステークホルダーと共創していくことが，オープンサイエンスに期待される役割である．クックパッドへのレシピ掲載にとどまらず，各種イベントなどの機会を活用し，実世界をフィールドとしたデータ活用に取り組む計画である．

4.3 アイデアソンの現実化

江戸料理レシピデータセットのアイデアは，2015 年 12 月に開催した「歴史的典籍オープンデータワークショップ〜古典をつかって何ができるか！じんもんそん 2015〜」から生まれたものである．筆者の一人（北本）はこのアイデアソンで，日本古典籍オープンデータの中に料理本が含まれることを初めて知った．そして参加者と議論するうちに，料理レシピをクックパッドに公開し，「つくれぽ」で料理法を報告し合えると面白いのではないかと考えた．また国立情報学研究所とクックパッドは，情報学研究データリポジトリにおける研究データ公開で以前から縁があった[6]．そこでアイデアソンの直後にクックパッドにコンタクトし，2016 年 1 月から議論を開始して，2016 年 11 月にはデータ公開に至った．このように江戸料理レシピデータセットは，アイデアソンを現実化したプロジェクトという面もある．

コンピュータビジョン研究で世界的に著名な研究者である金出武雄氏の言葉を借りれば，物事の実現には「素人発想，玄人実行」[7]が重要である．アイデアソンは，こうすれば使いやすいのにといった視点から素直にアイデアを出すという，いわば素人発想の場である．しかし素人発想のアイデアは，シンプルに見えてもいざそれを現実化しようとすると，多くの技術が必要になることが多い．つまりアイデアの現実化には，玄人の視点から価値あるアイデアを磨き上げ，専門的な知識に基づく設計と実行を進める段階も不可欠なのである．アイデアソンで盛り上がったアイデアを現実化するには，素人と玄人のコラボレーションが必須になると言える．

多くの人々が自由にアイデアを議論するイベントは確かに楽しいが，その場の楽しさだけで終わっては効果も限定的である．イベントで生まれた種をどう成長させて現実化するか，よい事例を共有できれば多くの人の助けになるだろう．

5．おわりに

本論文は，国文学研究資料館が中心となって推進する歴史的典籍のオープンデータ化を出発点とする，超学際的データプラットフォームの構築について論じた．最後にこの活動を推進する情報・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（http://codh.rois.ac.jp/）について紹介したい．このセンターは 2016 年 4 月 1 日に準備室としてスタートした新しい組織であり，2017 年には正式にセンターとして発足する予定である．このセンターでは，オープンサイエンスの潮流を踏まえたデータ駆動型の人文学を推進するために必要となる様々なデータやサービスを共同利用することを目指す．従来の人文学研究ではこうしたアプローチは少ないが，歴史的典籍 NW 事業のような大規模データをオープン化するプロジェクトも今後は増えるはずであり，その流れを後押しする情報基盤を提供することが，センターの重要なミッションであると考えている．

海外では HathiTrust や Google Books プロジェクトのように，網羅的な書籍デジタル化と OCR の成果を活用して，百年以上の長期間に及ぶ単語の頻度分布の変化から，文化の変化を探る研究なども進んでいる[8]．ところが日本でこうした研究が進展しないのは，書籍のデジタル化が遅れているだけでなく，機械による文字認識が難しいことも一因である．とはいえ，ビッグデータは文化を探るための新しい道具である．歴史的典籍 NW 事業のオープンデータが，当初計画通りの 30 万点やそれ以上の規模に到達すれば，日本文化の研究にも新しい可能性が開けるだろう．そこに機械学習という強力な武器で挑むという冒険に踏み出す研究者が増えることを期待する．

謝辞

IIIF Curation Viewer の作成には，フェリックス・スタイルの本間淳氏の協力を得た．江戸料理レシピデータセットの作成には，合同会社 AMANE の協力を得た．そして「クックパッド江戸ご飯」プロジェクトでは，クックパッド株式会社の伊尾木将之氏の協力を得た．

参考文献

1) 今西祐一郎: 大規模学術フロンティア促進事業「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」について，学術の動向，Vol. 21, No. 6, pp.6_66-6_68 (2016).
2) 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).
3) 服部正平：ヒトゲノム完全解読から「ヒト」理解へ，東洋書店 (2005).
4) 橋本雄太 ほか：くずし字学習支援アプリケーションの開発，研究報告人文科学とコンピュータ (CH)，2016-CH-110, pp.1-4 (2016).
5) 中野三敏：和本のすすめ，岩波新書 (2011).
6) 大山敬三，大須賀智子：情報学研究資源としてのデータセットの共同利用，人工知能学会誌，Vol. 31, No. 2, pp.254-261 (2016).
7) 金出武雄:独創はひらめかない—「素人発想，玄人実行」の法則, 日本経済新聞出版社(2012).
8) エレツ・エイデン，ジャン＝バティースト・ミシェル：カルチャロミクス　文化をビッグデータで計測する，草思社 (2016).