オープンサイエンスの動向と情報学分野へのインパクト

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

はじめに

PRMU研究会開催の趣旨

2016年4月から始まった第5期科学技術基本計画では「オープンサイエンスの推進」が明記され，研究のオープン化が重要な課題となっています．そこで，**研究のオープン化に関連する発表**を広く募集します．

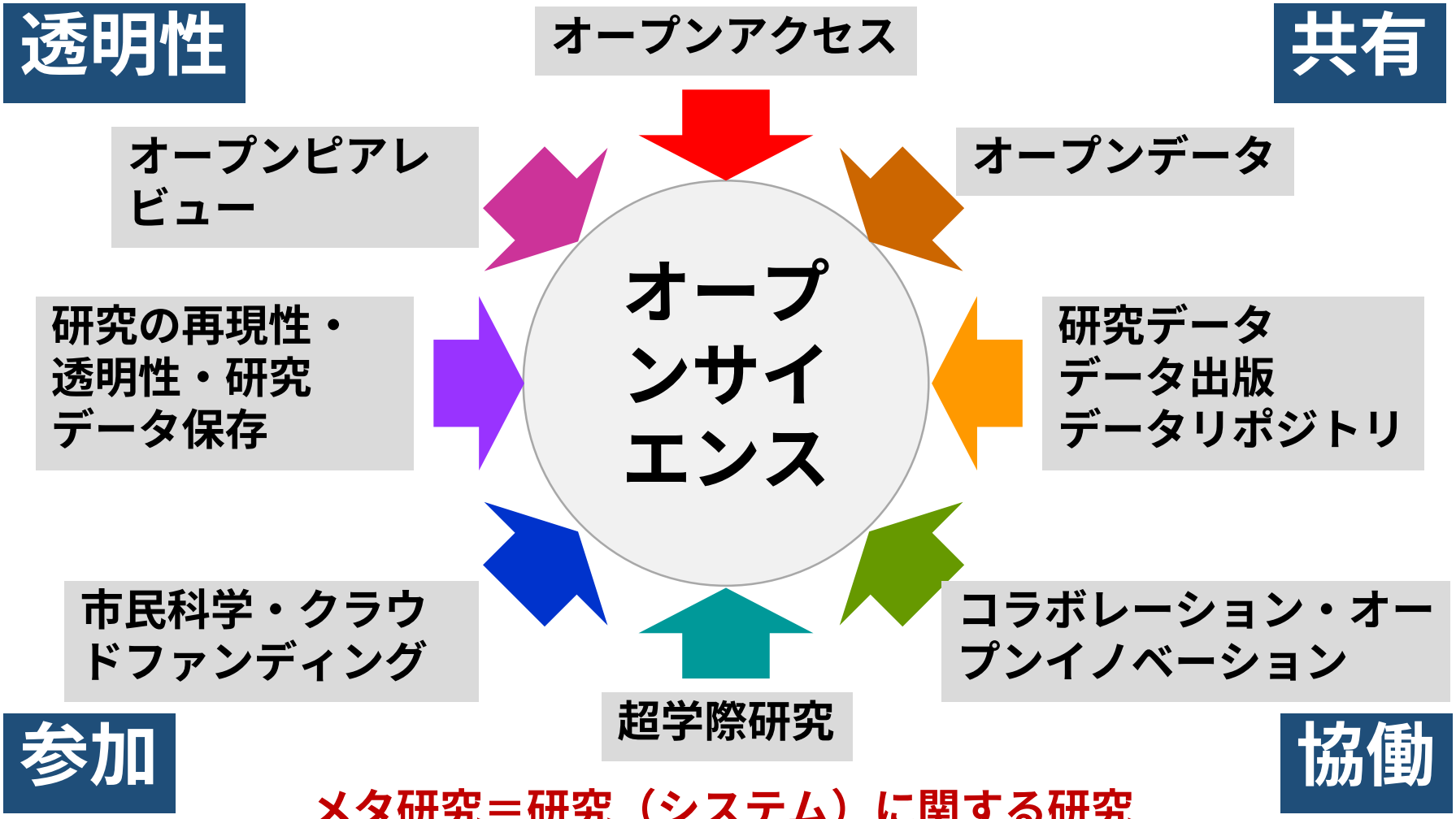
第一に，研究資源のオープン化にかかわる発表です．具体的には，研究用データセットや基盤的ソフトウェアをオープン化し，使いやすい形で提供することにかかわる発表です．例えば，データ品質向上のための前処理やメタデータ付与，アルゴリズムの工夫など，プロセスやノウハウに関する知見は従来の論文形式では共有しづらい傾向がありました．そこで，データベース（データセット）やソフトウェア自体を成果とみなすデータ論文・ソフトウェア論文を募集し，関連する知見を積極的にシェアしていきたいと考えています．

第二に，研究体制のオープン化にかかわる発表です．具体的には，同分野の研究者との共同研究にとどまらず，異分野の研究者や企業，専門家，市民などと協働してオープンなイノベーションを目指す研究などにかかわる発表です．例えば，人文学と連携するデジタル・ヒューマニティーズや，生物学と連携するバイオイメージ・インフォマティクスなど学術分野を越えた連携，そしてクラウドソーシングや市民科学など学術と社会が連携するタイプの研究は，従来の成果発表の場には収まりづらい傾向がありました．そこで，こうしたオープンな体制に基づく研究にかかわる論文を募集し，新しい可能性を積極的にシェアしていきたいと考えています．

オープンサイエンスとは？

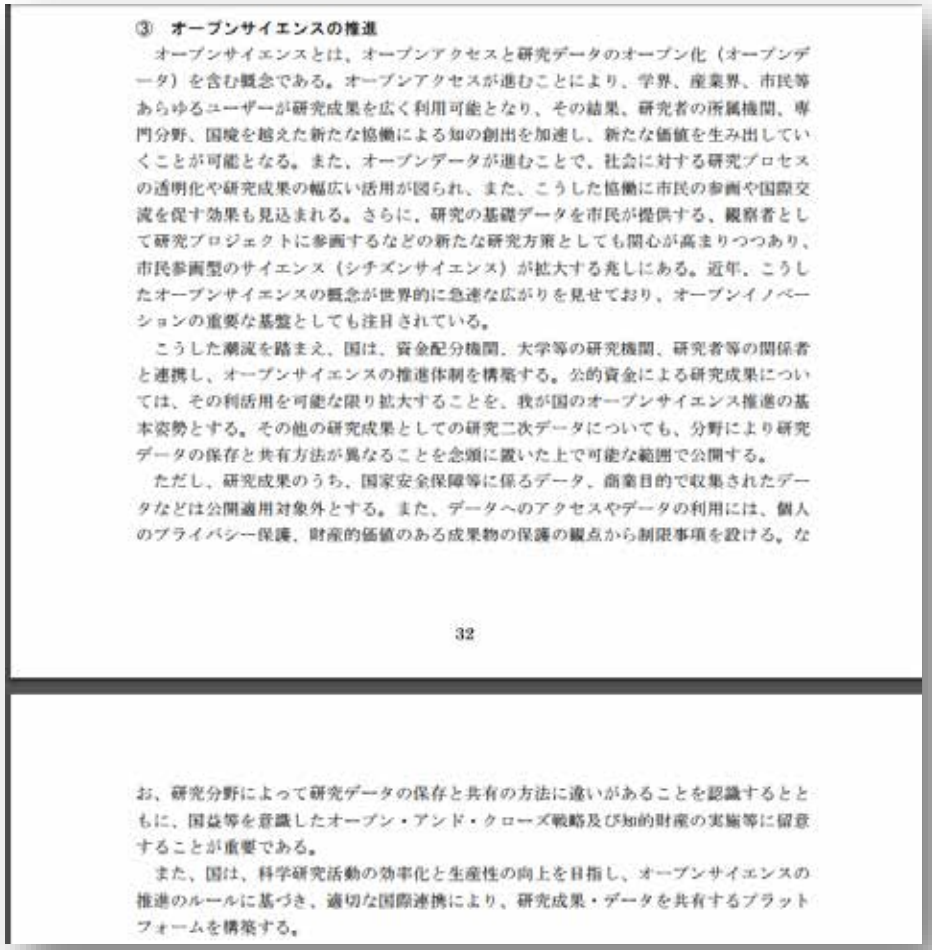
- 「オープン」という言葉を梃子にして、サイエンス（研究）の方向を変える。
- 「よりオープンに」という方向性を共有する活動を、一語で束ねると見える世界。
- 個々の活動ごとに「オープンサイエンス」の意味は異なり、単一の定義は困難。
- 大同団結？同床異夢？個々の活動を超える新しい目標を示せるかが問われる。

オープンサイエンスへの収束



メタ研究＝研究（システム）に関する研究

第5期科学技術基本計画



③ オープンサイエンスの推進

オープンサイエンスとは、オープンアクセスと研究データのオープン化（オープンデータ）を含む概念である。オープンアクセスが進むことにより、学界、産業界、市民等あらゆるユーザーが研究成果を広く利用可能となり、その結果、研究者の所属機関、専門分野、国境を越えた新たな協働による知の創出を加速し、新たな価値を生み出していくことが可能となる。また、オープンデータが進むことで、社会に対する研究プロセスの透明化や研究成果の幅広い活用が図られ、また、こうした協働に市民の参画や国際交流を促す効果も見込まれる。さらに、研究の基礎データを市民が提供する、観察者として研究プロジェクトに参画するなどの新たな研究方策としても関心が高まりつつあり、市民参画型のサイエンス（シチズンサイエンス）が拡大する兆しにある。近年、こうしたオープンサイエンスの概念が世界的に急速な広がりを見せており、オープンイノベーションの重要な基盤としても注目されている。

こうした潮流を踏まえ、国は、資金配分機関、大学等の研究機関、研究者等の関係者と連携し、オープンサイエンスの推進体制を構築する。公的資金による研究成果については、その利活用を可能な限り拡大することを、我が国のオープンサイエンス推進の基本姿勢とする。その他の研究成果としての研究二次データについても、分野により研究データの保存と共有方法が異なることを念頭に置いた上で可能な範囲で公開する。

ただし、研究成果のうち、国家安全保障等に係るデータ、商業目的で収集されたデータなどは公開適用対象外とする。また、データへのアクセスやデータの利用には、個人のプライバシー保護、財産的価値のある成果物の保護の観点から制限事項を設ける。な

32

お、研究分野によって研究データの保存と共有の方法に違いがあることを認識するとともに、国益等を意識したオープン・アンド・クローズ戦略及び知的財産の実施等に留意することが重要である。

また、国は、科学研究活動の効率化と生産性の向上を目指し、オープンサイエンスの推進のルールに基づき、適切な国際連携により、研究成果・データを共有するプラットフォームを構築する。

- 平成28年度～平成32年度の基本計画。
- **研究データの公開や活用、透明化、プラットフォーム**などが記述の中心。
- オープンアクセスや市民科学、オープンイノベーションなどにも触れる。

http://www8.cao.go.jp/cstp/kihonkeikaku/5honbun.pdf

PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE

- Recommendation 1: **Private and public institutions are encouraged to examine whether and how they can responsibly leverage AI and machine learning in ways that will benefit society**. Social justice and public policy institutions that do not typically engage with advanced technologies and data science in their work should consider partnerships with AI researchers and practitioners that can help apply AI tactics to the broad social problems these institutions already address in other ways.
- Recommendation 2: **Federal agencies should prioritize open training data and open data standards in AI**. The government should emphasize the release of datasets that enable the use of AI to address social challenges. Potential steps may include developing an "Open Data for AI" initiative with the objective of releasing a significant number of government data sets to accelerate AI research and galvanize the use of open data standards and best practices across government, academia, and the private sector.

https://www.whitehouse.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence

オープンサイエンスの背景

「オープン」の3つの側面

1. **他者が使える（再利用）**
   - オープンデータやオープンアクセスなど。外部の人が研究結果を自分の目的に再利用できる。
2. **他者が検証できる（透明性）**
   - オープンガバメントや研究再現性など。外部の人がエビデンスを検証し、正当性を判断できる。
3. **他者を受け入れる（参加）**
   - オープンイノベーションや市民科学など。外部の人を招きいれ、共に価値を生み出す。

オープンサイエンス革命
マイケル・ニールセン (著)
紀伊國屋書店, 2013
右の引用はp.340.

- **インターネットの活用で社会は変わった。では科学は？**
- 科学はオープン化を妨げる独自の問題に直面している。
- （CCなどのツールは）科学者の報酬は論文の出版によって得られるもの（中略）という問題
- オープンサイエンスは（中略）新たな考え方を必要としている。

ニールセンの「夢」を実現する方向に進んでいるわけでは必ずしもない。

データ駆動型サイエンス

1. **第四の科学**：データが核となるサイエンスへ。
2. **オープンなデータの必要性**：良質のデータがないと、研究やイノベーションが進まない。
3. **情報基盤**：データの生成、流通、分析、共有を支える新しい学術情報基盤が必要。
4. **インセンティブと人材育成**：持続的な活動にはインセンティブと人材育成の見直しが必要。
5. **超学際**：社会の課題解決に貢献するには、多様なステークホルダーが共創することが必要。

オープンイノベーション

- **知的財産のオープン化**：知的財産のオープン化が、協力者を「おびき寄せる」一つの戦略になった。
- **競争領域と協調領域**：差別化できる部分は守りつつ、外部の力を使えるところは使う。
- **コミュニティの形成**：参加者が増えれば、創出される価値も増える。
- **「壁」を超えた協調**：企業、学術、市民の間に新しい枠組みを作る。

オープンアクセスの顛末

オープンアクセス前	オープンアクセス後
• 論文の電子化をきっかけに購読料が高騰。 • 市民による論文アクセスが遮断される。 • グリーン（リポジトリ）、ゴールド（著者支払）で、論文アクセスをオープン化。	• 出版社は方針を変更し、著者支払いで十分な利益を得る。 • 何でも通す出版社のほうが利益が上がる。 • 詐欺的な出版社は、論文への永続的アクセスを放棄。

オープンサイエンスの駆動力

研究成果公開の歴史的変化

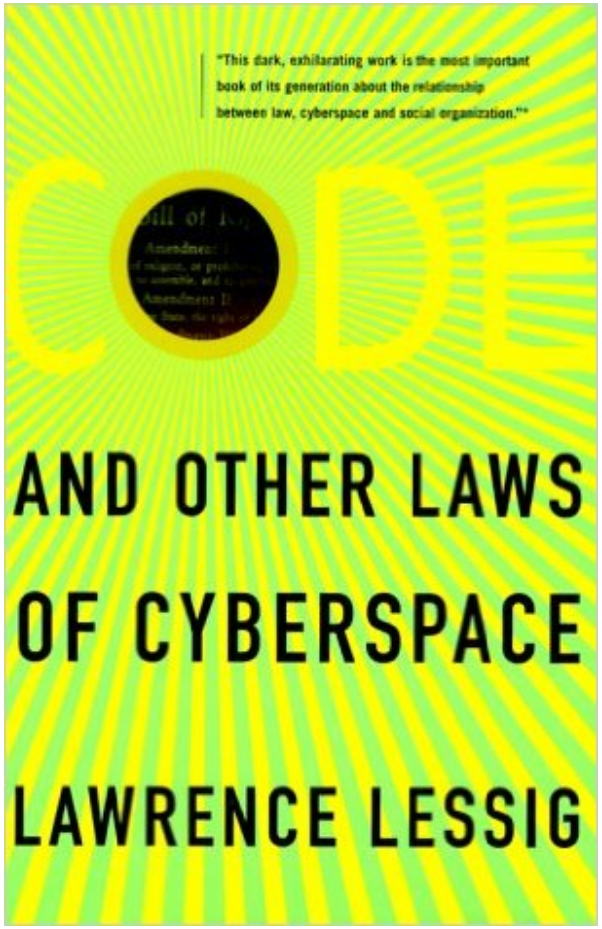
第一の変革期：17世紀後半

- **研究成果の秘匿から研究成果の公開へ。**
- Philosophical Transactions of the Royal Society (1665)
- **助成金を支給するパトロンが、名声のために成果公開を要求。**

第二の変革期：21世紀？

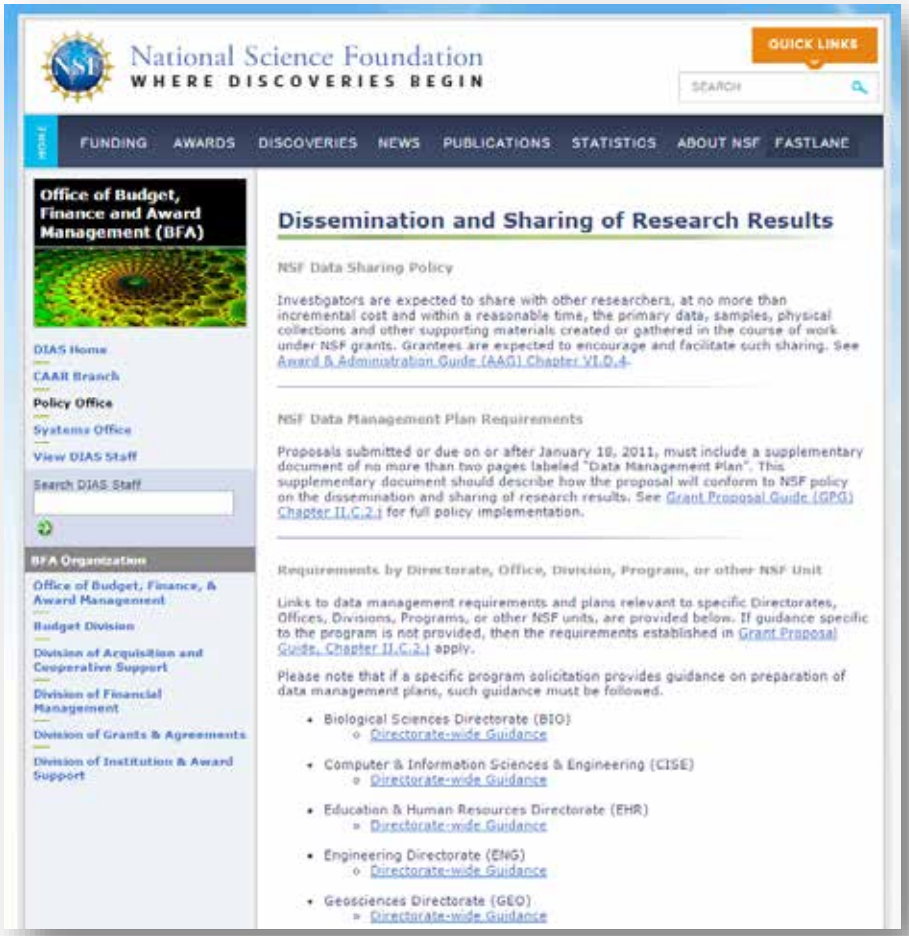
- **論文等の情報流通はほぼネットに移行。**
- **論文形式以外の研究成果も、簡単かつ即時に公開可能。**
- **資金提供機関が、公共利益のために成果公開を要求。**

制度を分析する4つの視点



- Lawrence Lessig (Founder of Creative Commons), *Code: And Other Laws of Cyber Space* (first edition 1999)
- **法＝しなければならない**
- **規範＝すべきである**
- **市場＝した方が利益がある**
- **アーキテクチャ＝せざるを得ない**

「法」によるオープン化



NSF Data Management Plan

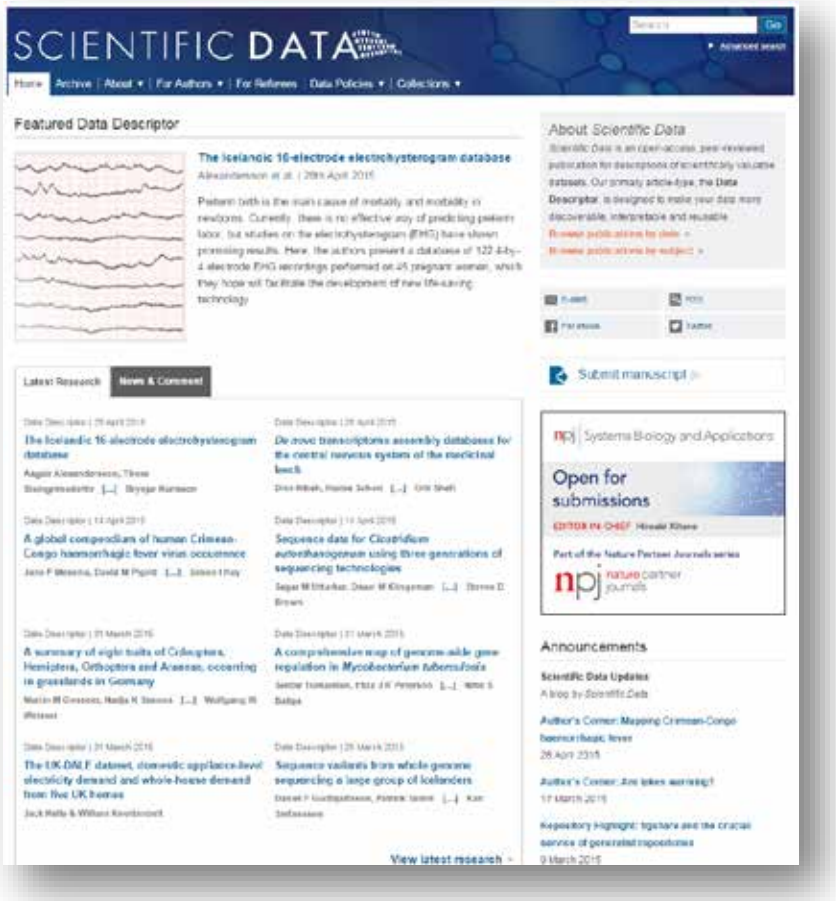
- **資金提供機関**が「データ管理計画」などを義務化。
- **研究評価**でも、オープン化の進展を指標に含める。
- 上からの押し付けには**副作用**も大きく、見極めが必要。

「規範」によるオープン化

https://www.icsu-wds.org/

- **オープンな文化**：データ共有が不可欠な分野もある。
- **世代の差**：若い世代では共有文化の経験がより強い。
- **文化の差**：異なる文化圏に対する説得力が弱い。

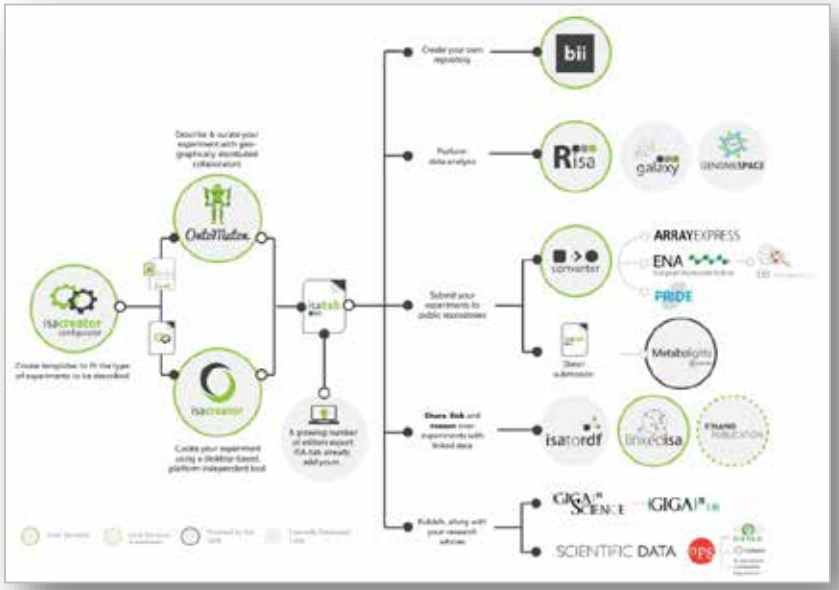
「市場」によるオープン化



Scientific Data (Nature publishing group)

- **報酬への期待**：研究成果をオープン化すると、引用も増加する [要出典]。
- **損失への不安**：他者に成果を横取りされるんじゃないの？報酬は労力に見合うの？

「アーキテクチャ」による オープン化



http://www.isa-tools.org/software-suite/

- **ルールの適用**：論文投稿するには、エビデンスデータもオープン化せねばならない。
- **苦痛の軽減**：有償サービスに任せた方が、オープン化が楽になる。
- **ベンダーロックイン**：良くも悪くも企業のビジネスチャンス。

オープンサイエンスと研究評価

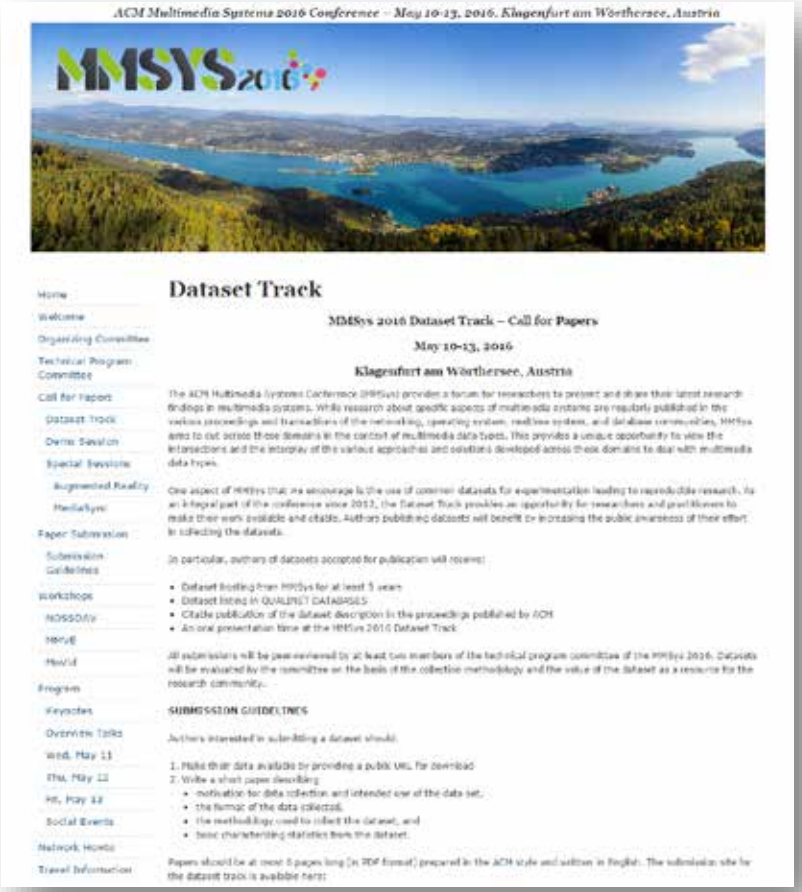
3種類の研究データ

研究資源データ	研究の入力となるデータ。評価用データセットなど。再利用のためのオープン化が求められる。
論文付属データ	研究の出力となるデータ。論文の図表やその元データなど。透明性のためのオープン化が求められる。
研究過程データ	研究の入力と出力の間で生み出されるデータ。日々の研究活動のエビデンスとなるデータなど。積極的にオープン化するものではないが、研究不正防止のための長期保存が求められる。

研究資源データセット

Olga Russakovsky*, Jia Deng*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. (* = equal contribution) ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015

データ論文



- **新規性より再利用性**：データセット公開と通常の研究成果では、評価基準が異なる。
- Data Descriptor：データセットの特徴を論文化する新フォーマット。
- **査読の導入**：高品質のドキュメントを残し、再利用性を高める。

https://mmsys2016.itec.aau.at/dataset-track/

データ引用

1. **ビブリオメトリクス**：引用（被引用数）は、昔から評価の最重要項目の一つである。
2. **データ引用**：論文向けの引用計測インフラに、データも相乗りできる（謝辞はできない）。
3. **データDOI**：データ（論文）に識別子（ID）を付与し、引用の計測を簡単にする。
4. **インセンティブ**：引用が増えて評価が高まれば、データ基盤の研究者もやる気が出る。
5. **持続性**：データ駆動型科学の持続性が高まる。

グローバルなIDと研究評価

1. **論文識別子**：DOI (Digital Object Identifier)は、論文識別子のデファクトスタンダード化。
2. **研究データ識別子**：DOIの付与が最近数年で徐々に普及、データ論文も増えつつある。
3. **研究者識別子**：ORCIDの普及が進むが、スタンダードとみなせるかは微妙な段階。
4. **ソフトウェア識別子**：GitHubへのDOI付与やソフトウェア論文など、まだ試行段階。
5. **研究助成識別子**：すべてのIDを結合すれば、プロジェクトの費用対効果も計算可能？

人文学オープンデータ共同利用センターにおける取り組み

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（国立情報学研究所の上位組織）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。
- **情報学＋統計学によるデータ駆動型人文学**で、人文学コミュニティの研究に新しい道を開く。

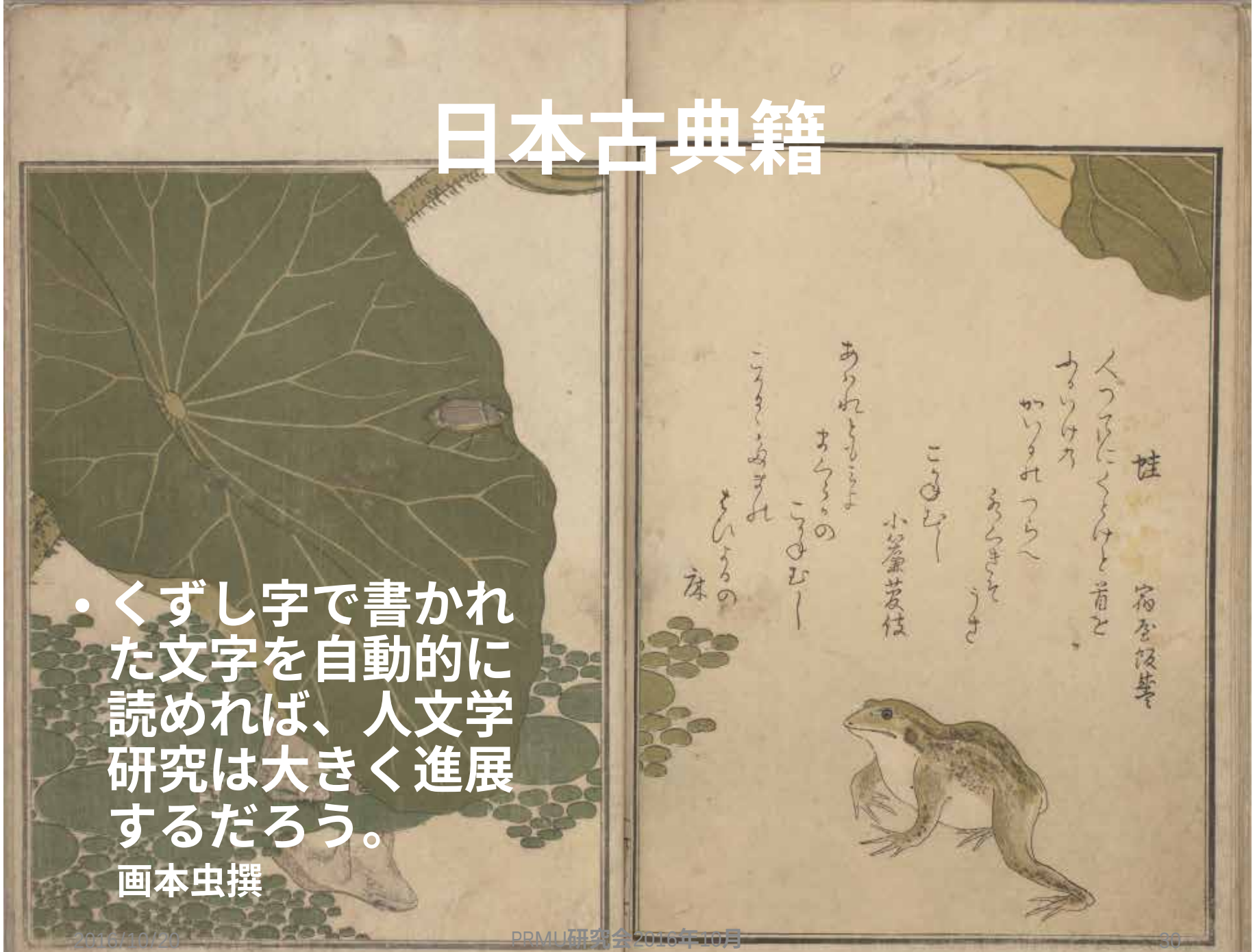
1. 情報学の技術を用いて人文学の研究を行う。
2. 人文学のデータを用いて情報学の研究を行う。

- **国文学研究資料館「歴史的典籍NW事業」**が、オープンデータの一つの中心となる。

国文研歴史的典籍NW事業

- 10年間で約30万冊の古典籍をデジタル化し、公開する（当初予算通りなら）。
- 国文研所有の古典籍は、350冊をCC- BY-SAで公開済み。近日中に700冊（158,455画像）に拡大。
- くずし字で書かれた版本を、一般人は読めない。翻刻（テキスト化）や現代語訳もまだ少ない。
- 人間が学習する＝モバイルアプリ（KuLA）
- 機械が学習する＝広義の文字認識＝ここはPRMUコミュニティが貢献できるところ！

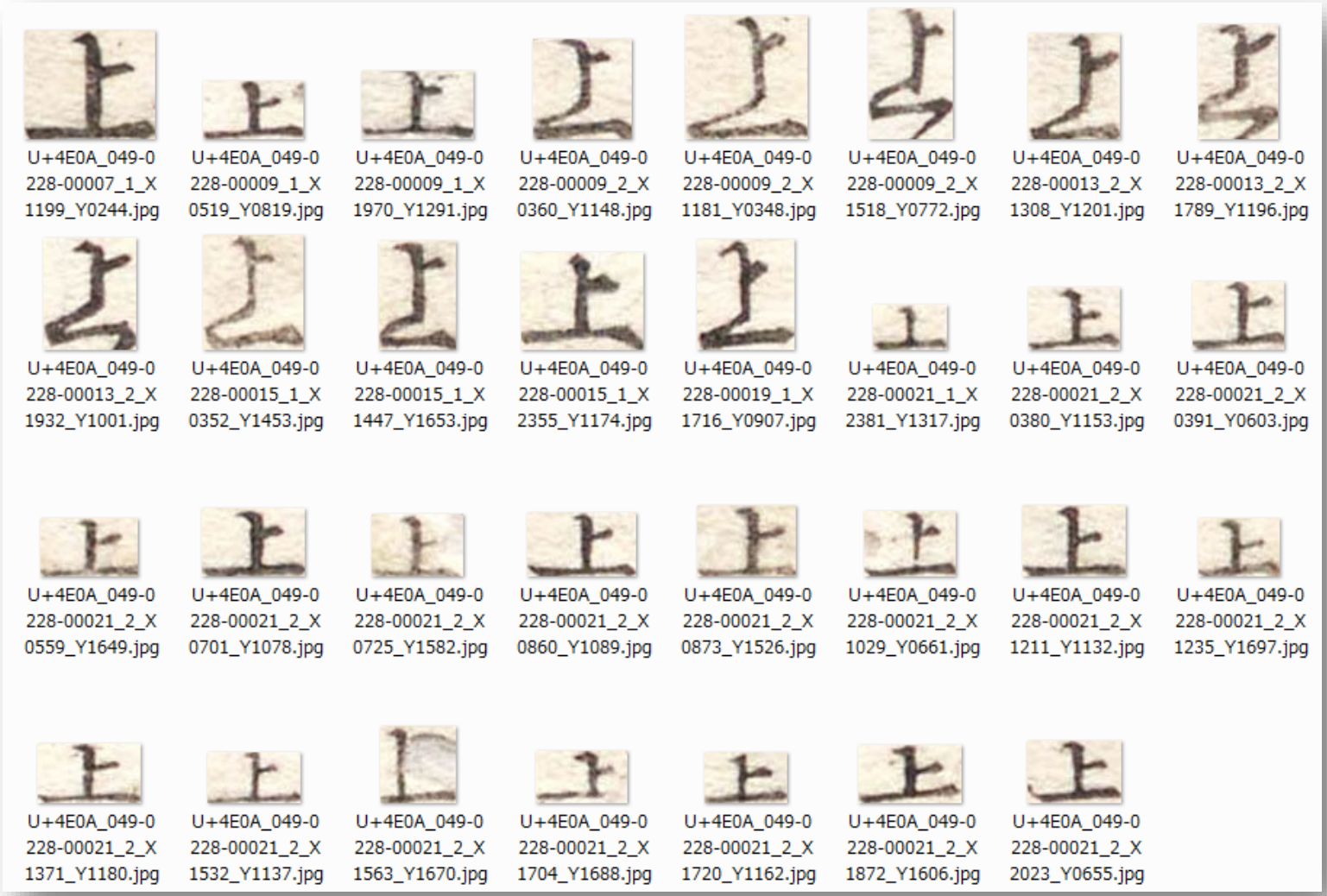
日本古典籍



- くずし字で書かれた文字を自動的に読めれば、人文学研究は大きく進展するだろう。

画本虫撰

字形データセット


U+4E0A_049-0228-00007_1_X1199_Y0244.jpg
U+4E0A_049-0228-00009_1_X0519_Y0819.jpg
U+4E0A_049-0228-00009_1_X1970_Y1291.jpg
U+4E0A_049-0228-00009_2_X0360_Y1148.jpg
U+4E0A_049-0228-00009_2_X1181_Y0348.jpg
U+4E0A_049-0228-00009_2_X1518_Y0772.jpg
U+4E0A_049-0228-00013_2_X1308_Y1201.jpg
U+4E0A_049-0228-00013_2_X1789_Y1196.jpg
U+4E0A_049-0228-00013_2_X1932_Y1001.jpg
U+4E0A_049-0228-00015_1_X0352_Y1453.jpg
U+4E0A_049-0228-00015_1_X1447_Y1653.jpg
U+4E0A_049-0228-00015_1_X2355_Y1174.jpg
U+4E0A_049-0228-00019_1_X1716_Y0907.jpg
U+4E0A_049-0228-00021_1_X2381_Y1317.jpg
U+4E0A_049-0228-00021_2_X0380_Y1153.jpg
U+4E0A_049-0228-00021_2_X0391_Y0603.jpg
U+4E0A_049-0228-00021_2_X0559_Y1649.jpg
U+4E0A_049-0228-00021_2_X0701_Y1078.jpg
U+4E0A_049-0228-00021_2_X0725_Y1582.jpg
U+4E0A_049-0228-00021_2_X0860_Y1089.jpg
U+4E0A_049-0228-00021_2_X0873_Y1526.jpg
U+4E0A_049-0228-00021_2_X1029_Y0661.jpg
U+4E0A_049-0228-00021_2_X1211_Y1132.jpg
U+4E0A_049-0228-00021_2_X1235_Y1697.jpg
U+4E0A_049-0228-00021_2_X1371_Y1180.jpg
U+4E0A_049-0228-00021_2_X1532_Y1137.jpg
U+4E0A_049-0228-00021_2_X1563_Y1670.jpg
U+4E0A_049-0228-00021_2_X1704_Y1688.jpg
U+4E0A_049-0228-00021_2_X1720_Y1162.jpg
U+4E0A_049-0228-00021_2_X1872_Y1606.jpg
U+4E0A_049-0228-00021_2_X2023_Y0655.jpg

字形データセットの公開


ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset

Overview

The "ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset" competition is organized in the framework of the ICFHR 2016 competitions by the Pattern Recognition and Human Language Technologies research centre with the collaboration of the READ partners. This contest aims to bring together researchers working on off-line Handwritten Text Recognition (HTR) and provide them a suitable benchmark to compare their techniques on the task of transcribing typical historical handwritten documents. Previous editions of this contest were organized at the ICFHR 2014 (Sánchez, 2014) and at the ICDAR 2015 (Sánchez, 2015).

The proposed dataset consists of a subset of documents from the Bentham collection composed of minutes of the council meetings held from 1470 to 1805 (about 10,000 pages), which will be used in the READ project. This dataset is written in Early Modern German, the number of writers is unknown. Handwriting in this collection is complex enough to challenge the HTR software.

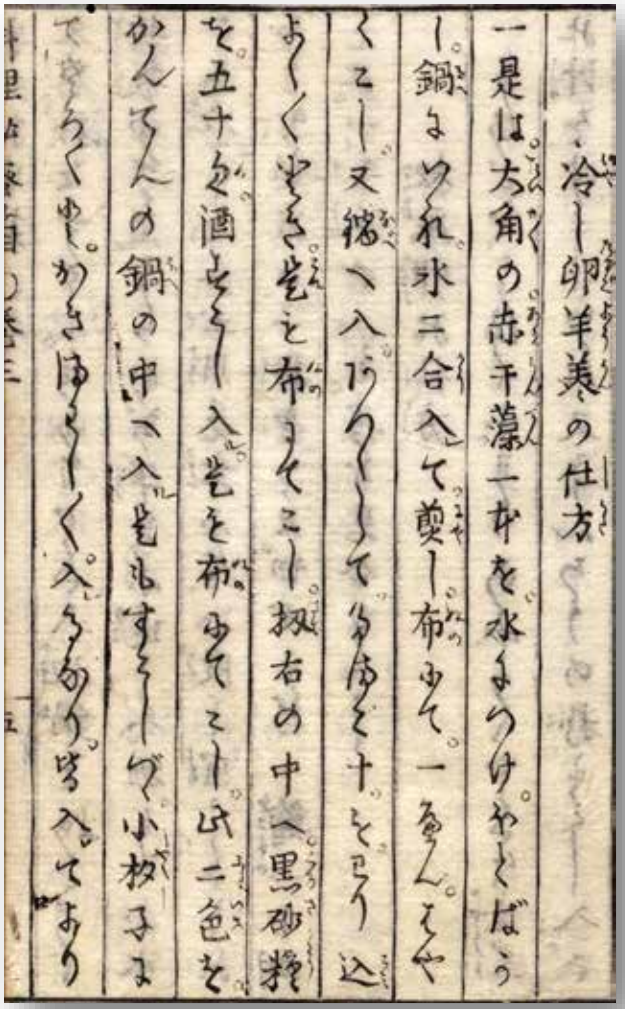
The dataset for this competition is composed of 400 pages; most of the pages consist of a single block with many difficulties for line detection and extraction (see page samples below). The dataset is divided into 2 batches for the competition: 1 batch for training and 1 batch for testing.

The first batch is composed of 400 pages. The ground truth of this set is in PAGE format (Pletschacher, 2010) and it will be provided annotated at line level in the PAGE files. For making easier the participation, several tools will be provided for extracting the lines as we describe below. Training data will be provided from March, 1st.

The second batch is a test set of 50 pages that will be kept hidden and released in due time just to obtain the results to be evaluated and compared.

- **字形データセット：第一弾となる約86,000文字を近日公開、今年度中に約40万文字に拡大。**
- 1) 基準画像、2) 基準画像上の座標（x,y,w,h）と文字コードの対応表、3) 上記座標で切り抜いた文字ごとの画像データ。
- **Historical Documentsを対象とした評価型ワークショップも今後は企画したい。**

超学際プラットフォーム



卵：10個
赤寒天（大角一本）：8g
葛粉
黒砂糖：約200g
酒：少し
水：360cc

研究者と市民が江戸レシピを共に探求

重要なお知らせ

- 人文学オープンデータ共同利用センターでは、近日中に人材募集（公募）を開始します。
- 人文学データに興味を持つ研究者の方々のご応募をお待ちしています！

おわりに

まとめ

- PRMUへのインパクト：PRMUはデータが重要な分野であり、研究データの観点からオープンサイエンスに取り組む価値がある。
- 革新性：人文学は最後のフロンティア？人間・文化に関する学問は、AI時代でも重要である。
- 持続性：研究コミュニティの核となるデータインフラの実現には、長期的な取り組みが必要。
- 詳しい情報はウェブサイトへ。
- http://codh.ex.nii.ac.jp/