NEWS RELEASE

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

平成 28 年（2016 年）11 月 17 日

江戸時代の文字の字形データセットを国文研との協働で構築
機械と人間の学習のためのオープンデータとして公開

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所（NII、所長：喜連川 優、東京都千代田区）は大学共同利用機関法人 人間文化研究機構 国文学研究資料館（国文研）(*1) と共同で、江戸時代の古典籍に書かれたくずし字の1文字ずつの字形画像データや文字座標データなどからなる「日本古典籍字形データセット」を制作し、11月17日から公開しました。本字形データセットは機械学習のための学習データセットとしての利用が期待されるとともに、人間のくずし字学習など教育目的への利用も考えられます。「日本古典籍字形データセット」は二次利用を歓迎するオープンデータ(*2) として、情報・システム研究機構の「人文学オープンデータ共同利用センター」準備室 (*3) のサイト（http://codh.rois.ac.jp/）から提供します。

今回公開するのは、江戸初期・寛文年間の料理本『料理秘伝抄』などの古典籍 8 点の画像データから 1 文字ずつ切り取ったくずし字 1,521 文字種の異なる字形のデータ計 8 万 6176 件です。字形のもとになった古典籍は、いずれも NII と国文研が共同で 11 月 10 日から公開を始めた「日本古典籍データセット」(*4) に収録されています。1 件の字形データは、以下の 4 種類のデータがセットになっています。提供する字形データは今後拡大し、今年度中に合計約 40 万件を公開する予定です。

1. **原本補正画像データ**
   翻刻作業を容易にするため、「日本古典籍データセット」で公開している画像に対して、見開き画像を分離し、回転させて正立させる前処理を加えた画像。
2. **文字座標データ**
   原本補正画像データ上で文字を取り囲む長方形の座標（XYWH）、文字の Unicode コードポイント、ブロック ID、文字 ID のデータ。
3. **字形画像データ**
   文字種ごとに字形を閲覧しやすくするため、「原本補正画像データ」に「文字座標データ」を適用して切り抜いた画像。
4. **作業報告文書**
   翻刻作業で読めなかった文字に関する情報やその他の注意事項を記載した文書。

National Institute of Informatics

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
総務部企画課 広報チーム
〒101-8430 千代田区一ツ橋 2-1-2
直通：03-4212-2164 FAX：03-4212-2150
E-Mail: media@nii.ac.jp

Web: http://www.nii.ac.jp
Twitter: @jouhouken
facebook: https://www.facebook.com/jouhouken

「日本古典籍データセット」のオープンデータ化により、古典籍に記された内容を画像で閲覧することが可能になりましたが、記された内容（文字情報、および、挿絵などの非文字情報）に対する検索は依然として大きな課題として残っています。「日本古典籍字形データセット」は、このうち文字情報の検索に関する研究開発に大きく寄与できると考えられています。具体的には、文字を画像から抽出して認識する OCR [*5] ソフトウェアの研究開発のための学習用データセットとしての利用が想定されます。本字形データセットを利用すれば、画像処理や自然言語処理、機械学習、人工知能などの分野の研究者がくずし字文字認識の研究に参画しやすくなるため、研究の活性化や研究者コミュニティーにおける知識共有への道が広がります。また、本字形データセットとともに、ディープラーニング [*6] を用いた文字認識プログラムをサンプルプログラムとして提供し、機械学習によるくずし字文字認識を気軽に試せるようにします。

「日本古典籍字形データセット」は、人間がくずし字を学習する目的にも使えます。本字形データセットを提供する「人文学オープンデータ共同利用センター」準備室のサイトでは、文字種ごとに文字のくずし方の違い（字形のバリエーション）を一覧することができます（別紙図）。例えば、平仮名の「し」の場合、一つ一つの文字の形の違いだけでなく、くずし字の元となった漢字（＝字母）「之」と「志」の違いによる異体字のバリエーションも画像で確認できるため、学習者は多くの字形を比較しながら学習を進めていくことができます。くずし字は日本語であるにもかかわらず、きちんと読める人は全国で数千人程度との推定 [*7] もあり、これは日本文化の継承における大きな問題です。本字形データセットは、問題解決に向けた情報学からの貢献にもなることが期待されています。

機械が賢くなるだけでなく人間も賢くなり、両者の力を合わせて日本古典籍の全貌を解き明かしていく――。NIIはそんな世界の実現に向けた研究開発を推進します。また、NIIは昨年11月に国文研との協働の第一弾として「国文研古典籍データセット」の公開に協力し、今月10日には共同で「日本古典籍データセット」をオープンデータとして公開しました。NIIは今後も組織の枠組みを超えて国内の研究機関のオープンデータ化を支援、推進し、日本の学術コミュニティーにおけるオープンサイエンス推進の担い手として、データ公開の流れが加速するよう取り組んでいきます。

【歴史的典籍オープンデータワークショップ（アイデアソン）】「日本古典籍字形データセット」の公開に合わせて、古典籍画像の利活用を考えるワークショップを12月9日に国文研と共催（予定）します。詳細や申し込みは、http://peatix.com/event/213615で。

National Institute of Informatics

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
総務部企画課 広報チーム
〒101-8430 千代田区一ツ橋 2-1-2
直通：03-4212-2164 FAX：03-4212-2150
E-Mail: media@nii.ac.jp

Web: http://www.nii.ac.jp
Twitter: @jouhouken
facebook: https://www.facebook.com/jouhouken

NEWS RELEASE

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

本件につきましては、本日、国文研も別途発表しています。国文研からは立川市政記者クラブに資料提供しています。国文研の発表内容は別紙の国文研プレスリリースをご参照下さい。

〈メディアの皆様からのお問い合わせ先〉

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所

総務部企画課 広報チーム

TEL:03-4212-2164　E-mail：media@nii.ac.jp

「日本古典籍字形データセット」「日本古典籍データセット」について

大学共同利用機関法人 人間文化研究機構 国文学研究資料館

古典籍共同研究事業センター事務室 古典籍共同研究係

TEL:050-5533-2988　E-mail：cijinfo@nijl.ac.jp

以上

（*1）国文学研究資料館：　国内各地の日本文学とその関連資料を大規模に集積し、日本文学をはじめとする様々な分野の研究者の利用に供するとともに、それらに基づく先進的な共同研究を推進する日本文学の基盤的な総合研究機関。http://www.nijl.ac.jp/

（*2）二次利用を歓迎するオープンデータとして提供：　利用条件は、クリエイティブ・コモンズ・ライセンス表示－継承4.0 CC　BY-SA（https://creativecommons.org/licenses/by-sa/4.0/deed.ja）。

（*3）「人文学オープンデータ共同利用センター」準備室：　平成28年4月1日、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に設置。NIIコンテンツ科学研究系准教授の北本朝展が室長となり、NIIと統計数理研究所との共同研究、および国内外の人文学研究機関との連携を軸に、人文学オープンデータの共同利用という課題に取り組む。来年4月にセンター化予定。

（*4）「日本古典籍データセット」：　古典籍700点の画像データ（約16万コマ）と書誌データのデータセット。作品紹介や翻刻テキストデータ、タグ情報が付いたものもある。詳細は、今年11月10日発行のニュースリリース「『日本古典籍データセット』公開で国文研と協働／国内研究機関のオープンデータの取り組みを支援・推進」（http://www.nii.ac.jp/userimg/press_20161110.pdf）参照。

（*5）OCR：　光学的文字認識（Optical Character Recognition）。デジタル画像上の文字を認識するソフトウェア。文字の切り出しを自動的に行い、画像を入力するだけで文字化できるソフトウェアもあり、文書の電子化などで活躍している。ただし、印刷文字に比べると手書き文字の認識は困難であり、くずし字OCRも開発されてはいるが、改良の余地も大きいのが現状。

（*6）ディープラーニング：　深層学習（Deep Learning）。多層のニューラルネットワークを用いた機械学習の方法。画像分類問題などにおいて従来方法を大幅に上回る性能を示し、その後も多くの分野で画期的な性能向上を示して大きな注目を集めている。文字認識でも活用は進むが、性能のさらなる向上には大規模な学習データが不可欠であるため、オープンなデータセットの公開へのニーズが世界的に高まっている。

（*7）「きちんと読める人は全国で数千人程度との推定」：　日本文学研究者（平成28年度文化勲章受章者）中野三敏による『和本のすすめ』岩波新書　（2011）参照。

〈図〉寛文年間の料理本『料理秘伝抄』で使われている文字種「し」の字形一覧画像

National Institute of Informatics

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
総務部企画課 広報チーム
〒101-8430 千代田区一ツ橋 2-1-2
直通：03-4212-2164 FAX：03-4212-2150
E-Mail: media@nii.ac.jp

Web: http://www.nii.ac.jp
Twitter: @jouhouken
facebook: https://www.facebook.com/jouhouken