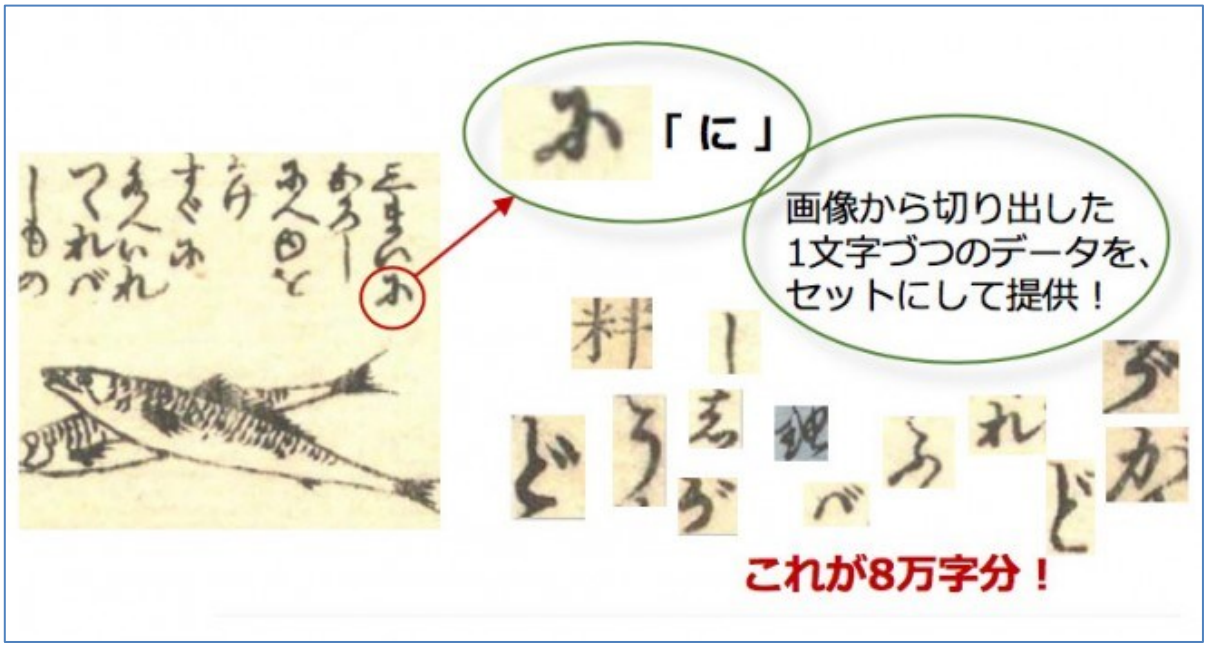
古文書を自動翻訳する日も近い！？　江戸時代の 8 万字超の「くずし字 字形データ」が無償公開へ

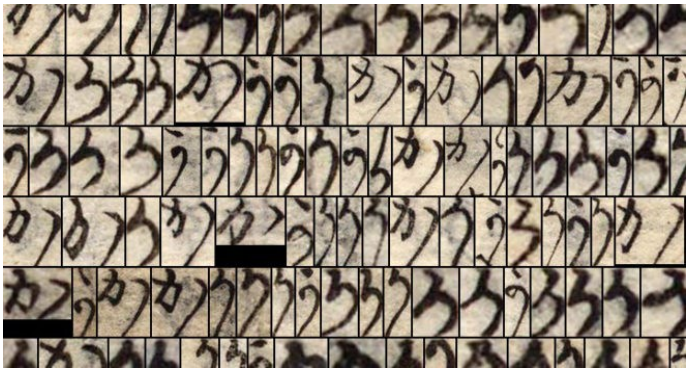


2016/11/18
大学共同利用機関法人 情報・システム研究機構 国立情報学研究所（NII）と大学共同利用機関法人 人間文化研究機構 国文学研究資料館（国文研）がすごいデータを無償公開しました！11 月 17 日に公開されたのは、**江戸時代の古典籍に書かれたくずし字の 1 文字ずつの字形画像データ。その数なんと 8 万 6176 件（1,521 文字種）**になります。

「どういうこと？」と思ってますか？ どれほどすごいことか、以下のリリース情報の画像解説を見ればわかります！
データは「日本古典籍字形データセット」という名称で、字形画像データのほか、文字が古典籍のどの位置に書かれているかを示す文字座標データと、原本の画像データも含まれています。
このデータは二次利用を歓迎するオープンデータとして無償提供中。

例えば、古典籍から抽出された「あ」のほんの一部。「あ」にも色々な癖がある。

例えば、古典籍から抽出された「か」のほんの一部。



収録されているくずし字は江戸時代に発刊された料理本 8 作品から 1 文字づつ切り出されたもの。料理本の一覧はこちら。

当世料理　万宝料理秘密箱　膳部料理抄　料理物語　日用惣菜俎不時珍客即席庖丁
料理方心得之事　新編異国料理　料理秘伝抄

「日本古典籍データセット」のオープンデータ化によって、画像から文字データを抽出する OCR ソフトや、機械学習、人工知能などの分野におけるくずし字認識システムを開発する上で、大きな役割を果たしてくれるのではないでしょうか。

このオープンデータを活用したシステムによって、近い将来江戸時代の古文書や浮世絵に描かれた文字などを、英語を日本語に翻訳するような感覚で自動翻訳できる日が来るのでしょうか。古典籍の中には解読されていないものもまだまだあるでしょうから、日本史の調査が加速するといった期待も持てますね。

くずし字をテキストデータ化する技術はすでに開発が進められていますが、今回のくずし字データベースの素材となるデータが無償提供されるというのは間違いなくビッグニュースですね。

字形データ数は今年度中に合計約 40 万字を公開する予定とのこと。