報道発表資料

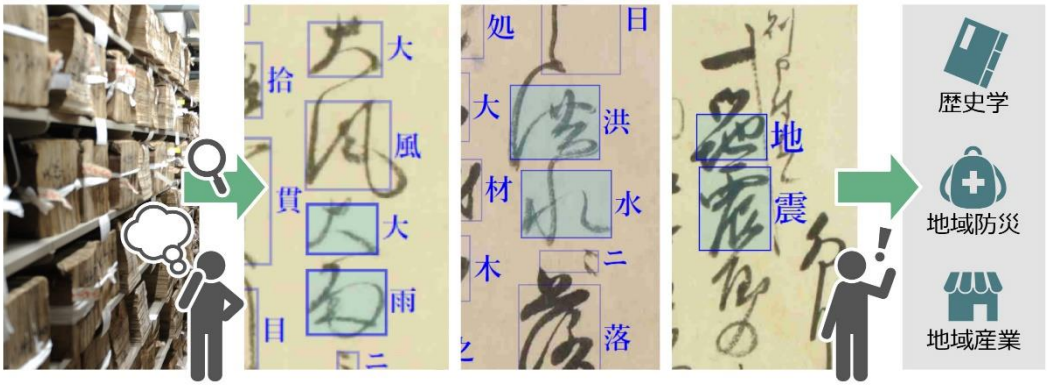
2024 年 7 月 26 日
国立大学法人 熊本大学
TOPPAN 株式会社

熊本大学と TOPPAN、くずし字 AI-OCR を活用した 古文書の大規模調査のための独自手法を開発

永青文庫所蔵の『細川家文書』における未解読の古文書約 5 万枚を、くずし字 AI-OCR を用い短期間で全文テキスト化に成功。文献資料の大規模調査のための独自手法を確立するとともに、新しく発見された災害関連の記録を用い、現代の防災計画への活用を目指す。

国立大学法人 熊本大学（所在地：熊本県熊本市、学長：小川久雄、以下熊本大学）と TOPPAN ホールディングスのグループ会社である TOPPAN 株式会社（本社：東京都文京区、代表取締役社長：齊藤昌典、以下 TOPPAN）は、熊本大学が公益財団法人永青文庫から寄託を受けている歴史資料『細川家文書（ほそかわけもんじょ）』のうち、専門家でも解読が困難な難易度の高いくずし字で書かれた約 5 万枚の未解読の古文書（藩政記録）を AI-OCR を用いて短期間で解読し、約 950 万文字のテキストデータを生成することに成功しました。

さらに、くずし字資料の解読システムと連動するキーワード検索システムを構築することにより、江戸時代前期の細川藩領国（小倉領 40 万石から熊本領 54 万石）の、約 90 年間にわたるあらゆる社会的事件や統治制度の変容を示す記述を含んだ資料を即時に検索収集できるようになりました。



大量のくずし字資料を AI-OCR でテキスト化し、検索可能にすることで様々な分野の研究等への活用が可能に

今回解読した古文書は、『細川家文書』のうち、細川家奉行所の執務記録である「奉行所日帳（ぶぎょうしょにっちょう）」、藩主細川忠利の口頭での命令を日次に記録した「奉書（ほうしょ）」、参勤中の細川藩主が国元の家老・奉行衆に発した書状の控えである「御国御書案文（おくにごしょあんもん）」、小倉・熊本の惣奉行衆から各業務を担当する奉行たちへ発せられた指示書類の控えである「方々（かたがた）への状控（じょうひかえ）」など、合計約 5 万枚です。

また、くずし字 AI-OCR により作成したテキストデータに対して、今回「地震、大雨、洪水、虫、飢、疫」などの災害に関連するキーワードで検索・調査を実施したところ 300 件以上の記述を発見しました。その中には、知られざる自然災害、疫病流行や飢饉など、歴史学・地域防災研究において重要な資料も含まれます。

今後、熊本大学と TOPPAN は、『細川家文書』の解読と分析を進め、江戸時代の社会史研究の通時的深化に貢献するとともに、新しく発見された災害関連の記録を活用することで、現代における防災意識の醸成、防災計画の策定等にも活用を目指します。

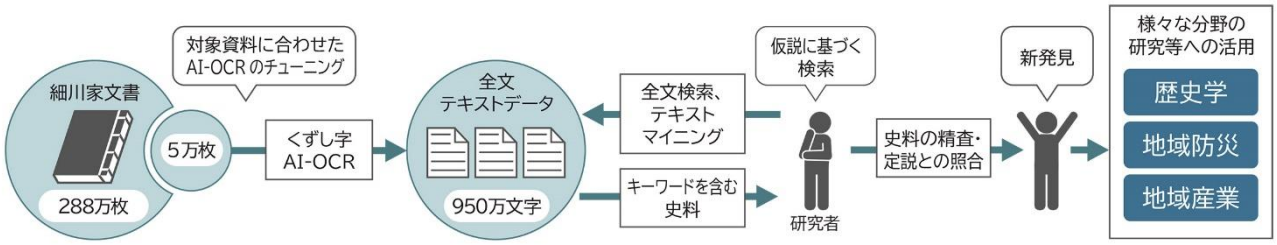
■ 取り組みの背景

古文書は、日本国内に数十億点以上残存すると言われていますが、その中には現代の社会課題にも直結する災害や地域文化の記録など、防災や観光資源の創出・地域の活性化にもつながる貴重な情報が記されているものがあります。しかし、古文書のほとんどは「くずし字」で書かれているため現代人にとって判読が困難となってしまい、当時の記録・文献を活用する際の大きな障壁になっています。

TOPPAN は、これらの課題を解決する新たな手法として、2015 年より大学共同利用機関法人人間文化研究機構 国文学研究資料館との共同研究を開始し、以後、多数の研究機関等とくずし字 AI-OCR 技術の開発・実証を重ねてきました。2017 年からは古文書解読とくずし字資料の利活用サービス「ふみのは®」として、様々なくずし字解読ソリューションを提供しています。

熊本大学は、公益財団法人永青文庫が所有する、九州の国持大名・肥後細川家（1600～1632 年 小倉藩主、以降 1871 年まで熊本藩主）に伝来した歴史資料や美術品のうち、約 5 万 7,000 点、約 288 万枚を寄託されています。寄託資料の中でも、今回解読した「奉行所日帳」をはじめとした、17 世紀初期から後期にかけて奉行所に蓄積された大量の統治記録は、当該時期の九州地域の社会状況を知る上でもきわめて貴重な歴史資料です。熊本大学では 2010 年に熊本大学永青文庫研究センターを設置し、永青文庫から寄託されている歴史資料や書籍等の基礎研究を推進して、ひろく社会に発信しています。

このような中熊本大学と TOPPAN は、2021 年より文献資料の新たな大規模調査手法の検討と、永青文庫所蔵資料に対する AI-OCR の精度向上に取り組んでおり、文部科学省における 2023 年度の科学研究費助成事業において『永青文庫資料と「くずし字 AI-OCR」の活用による 17 世紀社会論・公儀権力形成史の再構築』として採択されました。このたび約 5 万枚・約 950 万文字を全文テキスト化し、大規模な古文書解読のためのシステム構築を行うとともに、地域における災害記録をはじめとした網羅的な調査を開始しました。

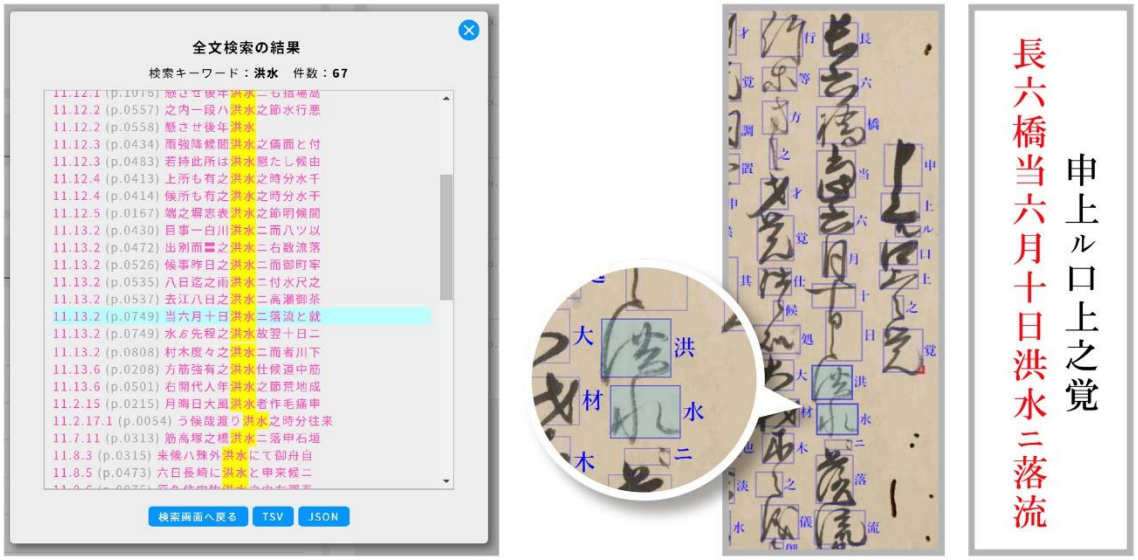


くずし字文献資料の大規模調査のフロー図

くずし字 AI-OCR による解読と検索システムが一体になることによって、これまでくずし字の解読が障壁となっていた古文書などの一次史料への網羅的調査が容易になります。検索により発見した資料を研究者が精査し、先行研究や定説との照合を行うことで、新たな発見や、歴史学をはじめとした様々な分野への一次史料の活用を促進します。

今回解読した『細川家文書』の約 5 万枚の資料に対し、災害に関するキーワード「大雨、虫、飢、疫」などで調査したところ、洪水、作物虫害、飢饉、疫病の発生と、それへの対応が行政課題化した事実を示す記述などを、300 件以上発見しました。

また、それらの中には、いままでよく知られていなかった 17 世紀後期の気象災害に起因する大規模な飢饉と疫病の蔓延を物語る熊本藩奉行所の執務記録の記述など、未知の重要な記述が含まれることが確認され、熊本における地域防災などに今後活用するための研究を進めていきます。



「奉行所日帳」に含まれる「洪水」の記述 67 件の中から
正徳2年（1712）旧暦 6 月 10 日の洪水で、熊本町の「長六橋」が流された記録を発見

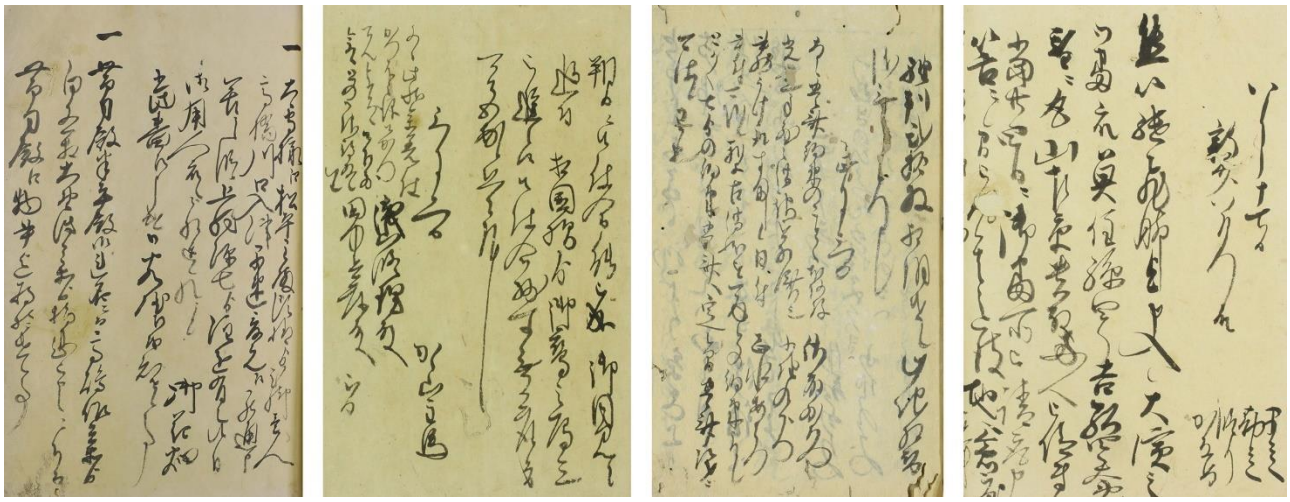
■ 両者の役割について

熊本大学：研究計画、史料選定、処理結果の評価、応用研究検討
TOPPAN：プロジェクトの管理・実行、くずし字 AI-OCR および解読システムの開発

■『細川家文書』について

『細川家文書』は、江戸時代に小倉藩主・熊本藩主をつとめた近世大名細川家に伝来した 5 万点以上、約 288 万枚の歴史資料群です。現在は公益財団法人永青文庫が所有し、その大半が熊本大学に寄託されています。『細川家文書』の資料群は主に以下の通りです。

① 「奉行所日帳（ぶぎょうしょにっちょう）」：小倉城や熊本城にあった細川家奉行所の日次の執務記録であり、当該時期の九州地域の社会状況が記された貴重な歴史資料
② 「奉書（ほうしょ）」：藩主細川忠利が、側近を通して口頭で命令した内容を日次で書きとめた記録
③ 「御国御書案文（おくにごしょあんもん）」：参勤中の細川忠利が国元の家老・奉行衆に発した書状の写し（案文）を集成したもの
④ 「方々（かたがた）への状控（じょうひかえ）」：小倉・熊本の惣奉行衆から各業務を担当する奉行たちへ発せられた書状の写しを集成したもの



左から、「奉行所日帳」、「奉書」、「御国御書案文」、「方々への状控」

■ 今後の展開について

今後、熊本大学および TOPPAN は、共同で『細川家文書』を解読し、当研究を通じて現代における防災計画や、歴史学の学習・研究の拡大に貢献します。

熊本大学は、『細川家文書』の解読と分析を進め、一時代の中でも細分化された短期間の枠内で完結するような研究法を克服して、江戸時代の長期にわたる社会変容の過程を通時的に把握し、九州に基点をすえた江戸時代社会史研究の深化に取り組んでいきます。

また、TOPPAN はグループ会社である TOPPAN デジタル、TOPPAN エッジとも連携し、AI-OCR による古文書解読支援システム「ふみのは®」の精度向上を目指すとともに、全国の様々な教育機関、博物館・資料館、地方自治体などと提携し、全国各地に眠る貴重な歴史的資料の研究・活用の支援に取り組んでいきます。

■ 熊本大学永青文庫研究センターについて

熊本大学永青文庫研究センターは、永青文庫資料をはじめとする熊本藩関係資料の総合的な研究を通じて当該資料に立脚した 拠点的研究を組織するとともに、文化行政機関等との連携によって地域文化振興に貢献し、もって熊本大学の教育、研究及び 社会貢献活動の充実発展に寄与することを目的として設立された研究組織です。

https://eisei.kumamoto-u.ac.jp/

■ TOPPAN グループについて

TOPPAN ホールディングスを持株会社とする TOPPAN グループは「印刷テクノロジー」をベースに「情報コミュニケーション」「生活・産業」「エレクトロニクス」の 3 分野で事業を展開しており、社会やお客さまの課題解決につながるトータルソリューションの提供を行っています。

・TOPPAN ホールディングス株式会社：https://www.holdings.toppan.com/ja/

・TOPPAN 株式会社：https://www.toppan.com/ja/

・「ふみのは®」について

「ふみのは®」は、古文書などのくずし字資料をデータ化し、解読をアシストする各種サービスの総称です。TOPPAN グループの培ってきた OCR 技術（自動文字認識）を活用し、くずし字の資料の解読や公開をサポートしています。

https://www.toppan.com/ja/joho/fuminoha/

・「くずし字 AI-OCR」技術について

OCR（Optical Character Recognition）とは光学文字認識のことで、文書画像に含まれる文字を読み取り、テキストデータに変換するソフトウェアの総称です。TOPPAN グループでは 2013 年から高い精度のテキストデータを提供する「高精度全文テキスト化サービス」を展開してきました。今回の取り組みで利用したくずし字 AI-OCR エンジンは TOPPAN デジタル株式会社にて独自に研究・開発しているもので、一つひとつの文字位置を認識できることを特長としています。そのため利用者はくずし字のような難読文字であっても、解読結果を古文書と照らし合わせながら読むことが可能です。

＜熊本大学教授、永青文庫研究センター　稲葉継陽センター長のコメント＞

世界史上の「近世」と呼ばれる時代において、日本ほど多くの文書が作成された地域は他にないといわれます。しかしこれまでは、独特の草書体で書かれた大量の近世文書から情報（キーワード）を引き出すには、それらを人力で解読して文字データ化する必要があり、必然的に膨大な時間を必要とするため、処理可能なデータ量が限られて、研究推進の大きな壁になっていました。

今次の「くずし字 AI-OCR」技術によって、この壁が取り払われ、文書の画像データさえ揃えれば、江戸時代の社会の長期にわたる変容の過程を把握することに道が開かれ、また長期にわたる地震や気象災害、さらに疫病（パンデミック）の発生傾向を瞬時に把握することが可能となります。

研究の飛躍的な発展が展望できるとともに、膨大な古文書の読み手が足りない地方文化財行政の現場にとっても、きわめて有益な新技術となるでしょう。

＜公益財団法人永青文庫　細川護光理事長のコメント＞

永青文庫には、大名細川家に蓄積された室町時代以来の史料が伝来しています。信長や秀吉をはじめとする天下人から受給した文書群は、すでに国重要文化財に指定されていて、ひろく知られていますが、近世藩政史料群は、なにぶん分量が膨大で、文字のくずし方も時代や書き手によって多様であるため、いまだごく一部分が利用されているに過ぎません。この「くずし字 AI-OCR」技術の発達には目をみはるものがあり、藩政史料の内容が学界さらに社会で本格的に活用される呼び水となることを期待します。

以　上

＜報道に関するお問い合わせ先＞
・熊本大学　総務部　総務課　広報戦略室
TEL:096-342-3269 / MAIL;sos-koho@jimu.kumamoto-u.ac.jp

・TOPPAN ホールディングス株式会社　広報部
TEL:03-3835-5636 / MAIL:kouhou@toppan.co.jp