人工知能（ＡＩ）技術の外交史料研究への利活用の探求
－－紙媒体史料の文字認識と課題－－

長岡 さくら

（日本経済大学経済学部経営法学科准教授／
福岡工業大学総合研究機構環境科学研究所客員研究員）

Sakura NAGAOKA（Associate Professor, Department of Management Law, Faculty of Economics, Japan University of Economics／
Visiting Fellow, Environments Research Laboratory, Comprehensive Research Organisation, Fukuoka Institute of Technology）

キーワード：ＡＩ技術、外交文書、文字認識

1. はじめに

現在、世界の国家の数は 196 ヶ国にのぼる[1]。世界に広がる其々の国家は、様々な外交活動を行う。国家による外交活動の様子は、当事国あるいは当事国以外の国家の外務省その他の機関により、何らかの形で文書が作成され、記録として残され保存されることが多い。例えば、我が国の外交史料館においては、令和元年 9 月現在、幕末から 1980 年代後半までの、外務省その他の機関が公的な目的で作成・保管してきた「外交史料」及び団体・個人から寄贈された「個人文書等」が所蔵されている。その構成内容は、幕末期の外交史料集である「通信全覧」「続通信全覧」の概要（正・続通信全覧）、明治期から第二次世界大戦終結までの外務省本省と在外公館のやりとりを記録した電報・公信、日本政府・外務省内の意志（原文ママ）決定に関する文書等を事件・事項別に整理した外交史料（戦前期「外務省記録」）、外交記録公開制度により公開された戦後期外交史料（戦後期外務省記録等）の目録、概要（戦後外交記録）など多岐に亘る[2]。

これらの外交史料は、国際法上、其々の国家による国家実行を知るための手段として役立つのみならず、国家の行動を分析する際にも非常に重要な手がかりとなる。なぜならば、各国家の外交活動や国家実行、そして、その背景や根拠は、必ずしもオープンになっている訳ではなく、ベールに包まれていることも多い。そして、当然のことながら、国家の外交活動や実行あるいはその記録は、一定の期間、秘匿にされることが多く、同時代的にこれを確認するのに困難を極めることが多い。従って、国際法研究において、国際法理論と国家実行の一致・不一致を確認し、様々な事象を体系的に検討するための素材・手段として、外交史料を、人工知能（ＡＩ）技術を用いて利活用することが可能となれば、より一層の国際法研究が進むものと考えられる。

2. 国際法研究における外交史料の利用の現状

さて、これまでの国際法研究においては、主として判例や二次文献を用いての研究が多かったように見受けられる。即ち、国際法理論についての検討が大きな比重を占め、国際法に関わる国際法主体の全ての活動である国際法実践自体や、国際法理論と国際法実践の間の齟齬等についてはほとんど検証されてこなかったように見受けられる。

勿論、国際法研究がこのような状況に至った理由の一つには、外交史料を用いた研究に存在する、その世界独特の難しさがあることが挙げられる。例えば、著者が執筆した博士論文においても同様の困難があった[3]。

著者は、博士論文において、国際法上の直線基線について、その成立過程とその後の各国による直線基線設定の国家実行、それに対する第三国の対応を題材として、国際法における抗議の実効性を分析した。同論文では、従来の国際法学では、著名な教科書は元より、外務省による広報説明等においても、国家は相手国（行為国）に対し抗議を行うことにより、自国は行為国による一方的行為の法的効果を排除しうる、あるいは、発現させないと、長い年月、その根拠を示さず記述され続けてきたことに着目した[4]。しかし、時々明らかにされていた米国政府の発言等を見る限り、これまでの国際法理論との不一致が見られるのではないかとの疑念が出てきた。これを確認するため、これまで従来の国際法研究ではほとんど行われてこなかった、各国の外交史料を中心とする一次資料に当たることにより、従来信じられてきた国際法理論の正しさあるいは誤りを見出そうとした。

そこで、当該問題に関する外交史料を中心とする一次資料を確認しようとしたところ、大きな問題に出喰わすこととなった。即ち、当該問題に関する外交史料がほとんど公開されておらず、見当たらなかったということに他ならない。同論文の構想及び執筆を開始した当初（2003 年）から

の数年間までは、同論文の執筆に必要となる我が国の外交史料において、その存在自体を確認できた史料は僅か 1 件にすぎなかった。また、当該史料についてもインターネットでは本文が公開されておらず、ファイルのタイトルが公開されているだけであった。そして、当該ファイルのタイトルから関連史料であろうことを推察できるに過ぎなかった。ファイルのタイトルがインターネット上で公開され始めたのも、インターネットが一般に広く普及し始めてからのことである。また、諸外国においても、米国政府が保有する一次資料及び各国が国際連合に提出した一次資料については僅かながらに公開されていたものの、国際法上の論点に関わる外交史料については、諸外国の公文書館のウェブサイトではほとんど存在自体が明らかとはなっていなかった。従って、国際法分野において外交史料を利活用した研究を行おうとする場合、伝統的に、史料の存否自体を自分の目で確認するために個々の研究者が各国の公文書館を訪れ、紙媒体の外交史料を一つ一つ確認する必要があった。そして、文書・史料の存在が確認できた場合、これらの写しを入手し、一つ一つ確認しながら分析する手法が用いられてきた。即ち、史料の確認及び収集自体、研究者個人の置かれている地理的な状況等に非常に依存する状態であったと言えよう。

このように、近年に至るまでは、外交史料の入手自体が困難を極めていたため、国際法研究において国際法実践／国家実行についての一次資料に依拠した国際法理論研究が行いにくかったということを指摘することができる。

しかし、近年、このような状況が徐々に変わりつつある。

例えば、我が国の外交文書については、1971（昭和 46）年 4 月 15 日の外交史料館の開館によって、ようやく一般の研究者が生の外交文書を閲覧可能な状態になったと言えるであろう。但し、1971 年の同館の開館から数年間は、いわゆる戦前期史料と呼ばれる幕末から第二次世界大戦終結までの外交記録に限定されていた。

戦後期の外交記録の公開が開始されたのは 1976（昭和 51）年のことである。但し、1976 年の第 1 回外交記録公開から 2008 年の第 21 回外交記録公開までは、その公開速度は非常に緩やかであった。1976 年から 2018（平成 30）年までの外交記録公開冊数の変化については以下の通りである（図 1 参照）。

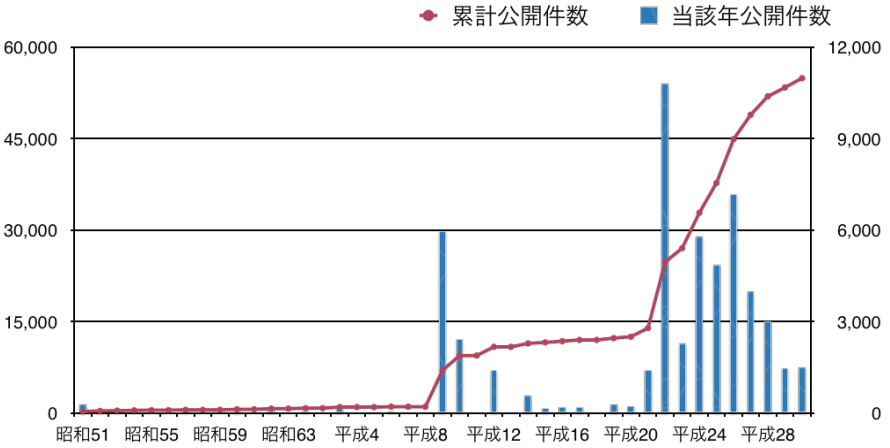


図 1　外交記録公開冊数の変化

出典：「外務省外交史料館所蔵史料検索システム」データに基づき著者作成

旧制度下の 1976 年から 2009（平成 21）年までの 24 年間に公開されたファイルは全部で僅か 13,958 冊に過ぎない。即ち、旧制度下の記録公開での年平均公開ファイル数は約 582 冊である。但し、この内、1997（平成 9）年から 1998（平成 10）年にかけて実施された第 13 回及び第 14 回の外交記録公開のみ、公開されたファイル数が突出している。また、旧制度下では、一冊も記録が公開されなかった年も存在する。当該二回分を除いた一回の平均公開ファイル数は、僅か約 218 冊に過ぎない。

従って、旧制度下の公開速度が続く限り、個々の研究者が一つ一つ目で文書を確認しつつ分析するという、上述したような伝統的な研究手法であっても、まだ十分に成り立つ状況であったと言えるであろう。

しかし、2010（平成 22）年以降、このような状況は著しく変化する。同年 5 月、外交記録公開の透明性を確保しつつ円滑に推進するために「外交公開に関する規則」が制定され、作成・取得から 30 年が経過した行政文書は公開するとの原則が明記された。また、2011（平成 23）年 4 月、公文書管理法が施行された。このように、公文書の管理等に関する法律の整備が行われたことによって、外交記録の公開が著しく加速することとなった。新制度下の 2010 年から 2018 年までの 9 年間に公開された外交文書は 40,973 冊にも上る[5]。新制度での公開開始時の一年間で 10,819 冊ものファイルが公開されたことを除いても、年平均公開冊数は約 5,122 冊に上り、現在もなお、少なくとも年間約 1,500 冊以上のペースで公開が進んでいるという状況にある。その結果、これまでに公開された 56,094 冊の全頁数は、一ファイルの頁数が数百頁に亘ることが多いことに鑑み、数千万頁

以上に上ると推定される。

勿論、このような状況は、国際法研究を行う者にとって大変利点があると考えられる。何故ならば、つい最近まで、そもそもどのような外交文書が存在しているかすら確認し難かったからである。

そもそも、我が国の外交文書のファイルタイトルの検索自体、利便性が増したのは2018（平成30）年末のことである。同年末までは、数ヶ月毎に新たに公開された文書名リストがＰＤＦファイルで公開されていたのみであった。同年12月14日、文書名及び各簿冊の小見出しが検索できる「外務省外交史料館所蔵史料検索システム」が外務省の公式ウェブサイトで公開されたことにより、多少の困難が解消されたものの、文書自体が公開されていないことには全く変化がない。即ち、秘密指定等が解除され、外交文書の存在自体が明らかとなったとしても、文書自体の入手には未だ困難が伴うと言えよう。

同時に、現在の状況には懸念も存在する。即ち、公開された厖大な外交文書を、一人の研究者が、特定の研究目的に従って、全ての関連ある資料を探し出し、読みこなし、分析を行い、それを一定の期間内に体系的な検討として公表することは今後可能なのかという疑問を抱かざるを得ない。

3. 紙媒体史料の文字認識と今後の課題

このような現状や困難を克服するため、本研究では、ＡＩ技術を用いて研究を一部補助させることを目指している。即ち、個々の研究者が行う外交史料の検討プロセスの一部を人工知能（ＡＩ）技術によって補助補完することができれば、ビッグデータとなった外交史料を駆使した研究が可能となるのではないかと思われる。

さて、ＡＩ技術を用い、研究を一部補助させるということは、これまでは、個々の研究者（人間）が行ってきた作業の一部を機械に代替させるということを意味する。それでは、人間が行ってきた作業を機械に代替させるためには何が必要となるであろうか。そのためには、研究プロセスを部分部分で分割する必要がある。研究者が研究を行う際、まず、文書を見る。そして、これを読み、理解し、文書について判断し、書く（執筆する）という一連のプロセスがある。本研究では、まず、文書を見て読むという流れ、次に、文書を読む（分析する）ことを機械に代替させることを目指している。

このため、まず、「文書を見て読む」ことを機械に代替させる作業に取り掛かった。「文書を見て読む」ためには、文書がどのような言語・文字で書かれているかを把握する必要がある。そのため、コンピューターが利用できる文字コード（文字データ）化されていない文字――いわゆる、手書き文字や文字画像――については、これをイメージスキャナやドキュメントスキャナを用いて読み取り、文字データに変換する光学的文字認識（Optical Character Reader：ＯＣＲ）作業を行う必要がある。

さて、近年、情報関連技術の発達により、法学分野においても実に大量のデータが見られるようになってきた。しかし、法学分野ではまだ紙媒体で提供される資料が大多数であると言える。たとえ、デジタル化された資料があったとしても、その多くは、ドキュメントスキャナ等を用いた、紙媒体を文字画像化した資料であり、先に述べたような文字データ化された資料はまだ少数であると言えよう[6]。国際法分野においても、近年、欧米諸国で発行された書籍や雑誌の一部は、紙媒体文献と文字データ化された文献の双方が発刊されているが、我が国においては、ほとんどが紙媒体のみで発行されている。

本研究では、まず、これまで著者が収集し検討を重ねてきた国際海洋法分野の外交史料や国会議事録等をサンプルデータとして用い、現存するＡＩ技術の到達度を明らかにする作業・検討に取り掛かった。その第一段階として、紙媒体あるいは文字データされていない画像データ（文字画像）について、市販されている汎用的なＯＣＲソフトを用いて文字画像の認識作業を行った。

現在、外交記録公開制度にて提供されている外交史料には、主として以下の二つの特徴がある。

第一に、提供される史料媒体についてである。現在の制度下では、大別して史料の原本、マイクロフィルム、ＣＤ－Ｒといった媒体による閲覧が可能である。なお、これらの史料のうち、マイクロフィルムやＣＤ－Ｒにて提供される史料はいわゆる文字画像であって、文字データ化された史料ではない。また、原本にて供される史料は紙複写及びスキャナ等での撮影及びＣＤ－Ｒ等への書込み、マイクロフィルムやＣＤ－Ｒにて供される史料は同形式媒体への複写及び紙出力を依頼することが可能である（有料）[7]。

第二に、史料の文字形態についてである。一定の時代・時期までの史料は、そのほとんどが手書き文字で構成されている。また、電報（案）の決裁過程において、更にこれを手書きにて線入れ・修正を行なっていることから、単に整然と文字が並んでいるという状態ではないという特徴がある。

これらの史料の、ＯＣＲソフトを用いての文字画像の認識作業にあたっては、文系研究者が利用しやすいと思われるソフト及び比較材料として高精度の日本語認識エンジンを搭載していると予想されたソフトとして、次の三つを使用して認識状態の確認を行なった。なお、認識文字の設定が可能なものについては日本語設定を行った上で認識作業を行った。

① Google ドライブ
② 読取革命 Ver.15（パナソニックソリューションテクノロジー株式会社）
③ WinReader PRO v.15（株式会社ＮＴＴデータＮＪＫ）

サンプル画像として以下の画像データを使用した、其々のソフトの認識結果は次の通りである。

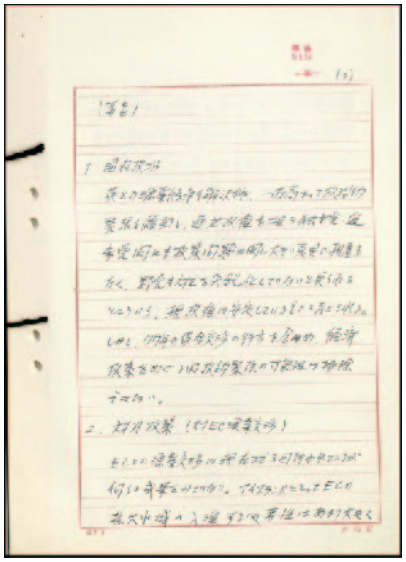
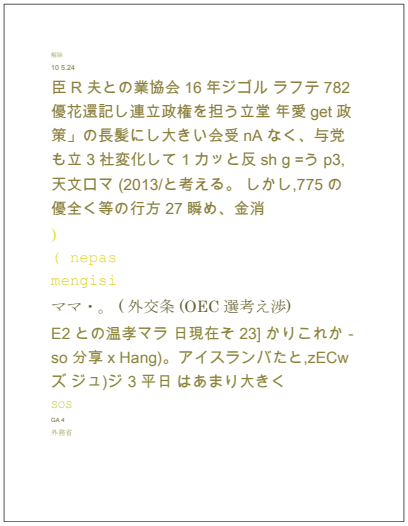


図２　ＯＣＲ作業に用いたサンプル画像[8]



10 5.24
臣 R 夫との業協会 16 年ジゴル ラフテ 782 優花還記し連立政権を担う立堂 年愛 get 政策」の長髪にし大きい会受 nA なく、与党も立 3 社変化して 1 カッと反 sh g =う p3, 天文口マ (2013/と考える。 しかし,775 の優全く等の行方 27 瞬め、金消
)
(nepas
mengisi
ママ・。 (外交条 (OEC 選考え渉)
E2 との温孝マラ 日現在そ 23] かりこれか - so 分享 x Hang)。アイスランバたと,zECw ズ ジュ)ジ 3 平日 はあまり大きく
sos

図３　「Google ドライブ」による認識結果

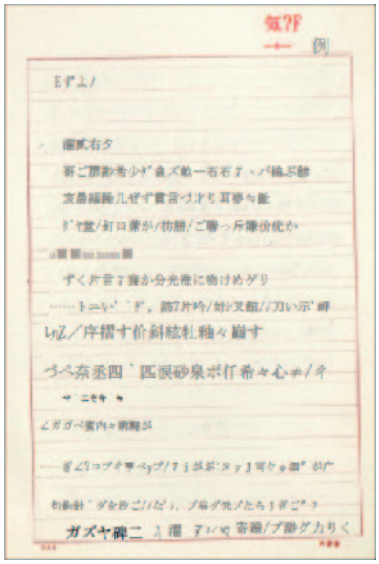


図４　「読取革命」による認識結果

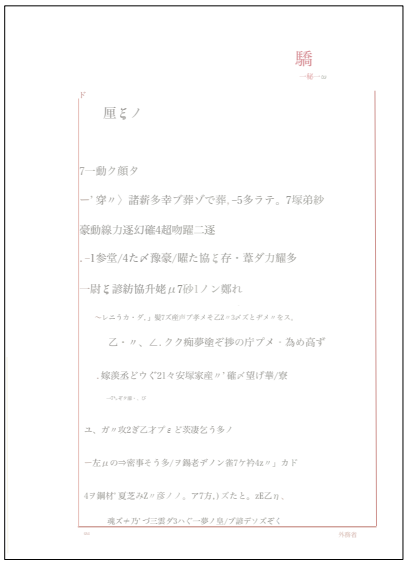


図５　「WinReader Pro」による認識結果

当該作業の結果、使用したサンプルデータに代表されるような外交史料については、言語の種類、文字の態様、文書の形式、綴り込まれている史料の多様性等の問題から、汎用的なＯＣＲソフトでは対応できないことが判明した。

本研究では、当初、紙媒体で提供され、これを画像化した一連の外交史料を主たるサンプルデータとして用い、研究を進める予定であった。ところが、研究を進めるにつれ、文字データとして提供されていない一連の外交史料は、言語の種類、文字の態様、文書の形式、綴り込まれている史料の多様性等の問題から、市販されている汎用的なＯＣＲソフトを用いた簡便な文字データ化には、ソフトの調整だけでは越えられない課題があることが判明した。その際、複数のＯＣＲソフトを用いて確認を行ったが、一データ中に複数の言語等が混在する外交史料についてＯＣＲソフトでは文字データ化すら十分にこなせないことが判明した。

このため、続いて、この技術的課題を回避するための作業を中心に進めた。そこで、画像化された史料のＯＣＲ作業に関しては、パターン認識による文字認識アルゴリズム自体を構築可能な数値解析ソフトであるＭＡＴＬＡＢに変更して対応することとした。

ＭＡＴＬＡＢを用いての文字認識アルゴリズムの構築には、プログラミング環境を整備する必要があったため、新たに当該環境を整えた上で作業を実施した[9]。

なお、ＭＡＴＬＡＢを用いたＯＣＲ作業においては、Computer Vision Toolbox に付属するＯＣＲトレーナーアプリにて、ＯＣＲにて認識できなかった文字を学習することが可能である。このアプリでは、学習の際に対話的に文字データにラベル付けを行い、関数 ocr で使用するＯＣＲ言語データファイルを生成できる[10]。

このため、ＭＡＴＬＡＢを用いたＯＣＲ作業にて認識できなかった文字については、次の工程として、ＯＣＲトレーナーアプリにてテキストイメージをセグメント化し、文字を学習させる作業を行なっている。

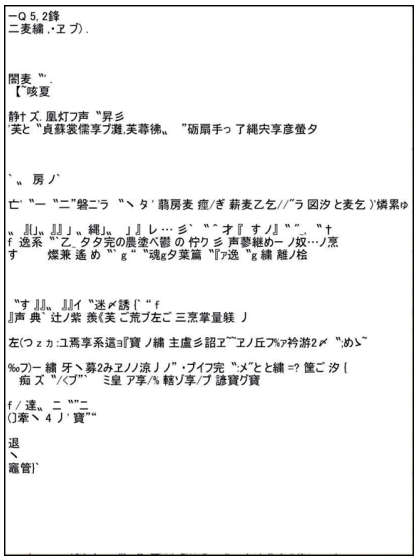


図６ ＭＡＴＬＡＢによる当初の認識結果

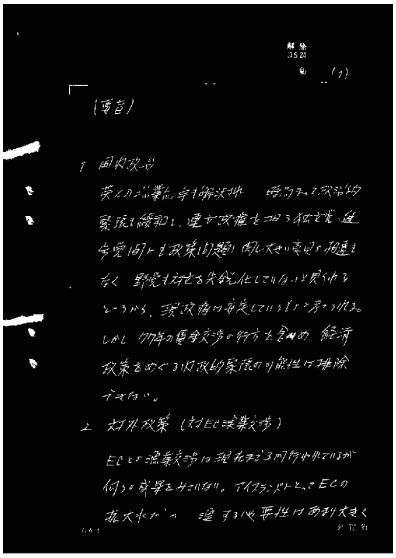


図６ ＭＡＴＬＡＢ・ＯＣＲトレーナーアプリを用いた図２テキストイメージのセグメント化

その結果、ＡＩ技術を用いて、研究者がこれまでに人力で行ってきた資料の「探し出し、読みこなし、分析する」といった一連の作業の一部を代替させることでさえ現状では課題が多いことが判明した。

また、これまでの検討の結果、主として次の点が判明した。即ち、文字データ化された文書の解析については、ＱＤＡソフトによる解析は、ＡＩ技術の開発（実装）についての研究として本研究を展開させるには適していないことが判明した。そこで、本研究を更に進めるためには、既に文字認識アルゴリズムを構築した数値解析ソフト環境を用いて、新たにデータマイニングのアルゴリズム構築に取り組むことが必要であると考えている。これにより、既に開始した文字データ化された文書の解析作業を文字認識アルゴリズムと一体的に進めることが可能になるのではないかと考える。また、データマイニングのアルゴリズム構築にあたっては、他分野における先行事例を参考に、外交史料研究において利活用可能な手法等について更なる検討が必要であると考えている。

現在、これらの観点から、数値解析ソフト環境を用いた、データマイニングのアルゴリズム構築に向けた検討を進めている。これらの検討結果については、数値解析ソフト環境を用いた文字認識アルゴリズムの構築についての検討結果と併せて、別稿に譲ることとしたい。

---

[1] 外務省「世界と日本のデータを見る（世界の国の数、国連加盟国数、日本の大使館数など）」（平成 31 年 3 月 29 日）、https://www.mofa.go.jp/mofaj/area/world.html、参照（2019 年 11 月 20 日最終確認）。なお、当該数は日本が国家承認を行なっている国家数を示したものである。従って、例えば、北朝鮮（朝鮮民主主義人民共和国）のように、日本が国家承認を行なっていない国は含まれていない。

[2] 外務省「外交史料館：所蔵史料の概要」（令和元年 9 月 20 日）、https://www.mofa.go.jp/mofaj/annai/honsho/shiryo/shozo/index.html、参照（2019 年 11 月 20 日最終確認）。

[3] 長岡さくら、『国際法における直線基線の設定と第三国の対応――国家実行における抗議の実効性――』（九州大学提出、博士学位論文、2016 年）。

[4] 例えば、外務省による、竹島問題について広報説明を行なっているパンフレットでは、大韓民国による竹島の不法占拠について次のように説明している。「・・・＜略＞・・・、韓国は、いわゆる『李承晩ライン』を一方的に設定し、そのライン内に竹島を取り込みました。これは明らかに国際法に反した行為であり、我が国として認められるものではない旨、直ちに厳重な抗議を行いました。・・・＜略＞・・・。このような韓国の力による竹島の占拠は、国際法上一切根拠のないものであり、我が国は、韓国に対してその都度、厳重な抗議を行うとともに、その撤回を求めてきています。こうした不法占拠に基づいたいかなる措置も法的な正当性を有するものではなく、また領有権の根拠となる何らの法的効果を生じさせるものでもありません。」「国際法に反した李承晩ラインの一方的設定により日本との領有権紛争が発生した後に、韓国が日本の一貫した抗議を受ける中で行っている一連の行為は、国際法上、証拠力が否定され領有権の決定に影響を与えることはありません。」日本国外務省、『竹島 なぜ日本の領土なのかがハッキリわかる！Ｑ＆Ａ付き竹島問題１０のポイント』（外務省アジア大洋州局北東アジア課、2014 年）、とりわけ、4 頁。

[5] 2018 年末までの公開冊数は計 54,931 冊。2019 年 6 月 15 日現在、全公開冊数は 56,094 冊。

[6] 近年では、汎用的なドキュメントスキャナの普及により、文系の研究者であっても紙媒体資料の文字画像化までは簡便に行うことが可能である。

[7] 外務省外交史料館利用細則第 4 条に係る複写については、以下に料金表が掲載されている。
https://www.mofa.go.jp/mofaj/files/000404179.pdf、参照（2019 年 11 月 20 日最終確認）。

[8] 在スウェーデン都倉大使発外務大臣宛交信秘第 45 号「管内情勢報告の提出（アイスランド）」昭和 52 年 1 月 20 日発、外務省記録「アイスランド・英国紛争（タラ戦争）」、2010-3597/SA.1.6.4、外務省外交史料館、より抜粋。

[9] なお、我が国の外交文書を用いたＯＣＲ作業の後、文字画像認識に用いるサンプルデータの見直しも行い、別途、作業を行った。当該作業過程においては、書体や言語等の混在しない１９世紀に公刊された資料の画像データを利用するとともに、念のため、以前に用いたＯＣＲソフトによる文字データ化の作業結果とＭＡＴＬＡＢを用いて作成した文字認識アルゴリズムによる文字データ化の作業結果との比較も行った。本作業結果については、別稿にて述べる予定である。

[10] ＭＡＴＬＡＢドキュメンテーション「カスタム フォントについての光学式文字認識の学習」、
https://jp.mathworks.com/help/vision/ug/train-optical-character-recognition-for-custom-fonts.html、参照（2019 年 11 月 20 日最終確認）。

謝辞：本研究は JSPS 科研費 17K18549 の助成を受けたものである。

（令和元年１１月２０日受付）