防衛省防衛研究所仕様書			
件　名	戦史史料のＯＣＲ学習データ等の作成	作　成	戦史研究センター

１．適用範囲

この仕様書は、防衛省防衛研究所が所蔵する戦史史料の利活用に資する学習用データ等の作成について規定する。

２．役務に関する要求

２．１　作業内容

契約相手方は以下の作業を実施する。

同所が保持する旧陸海軍文書の画像データを作業側に貸与する。(官側は当該画像データを格納したハードディスクを準備、貸与する。)

（１）基礎史料調査

防衛研究所戦史研究センター史料室が保有する史料について、現地（当研究所Ｆ１棟史料庫）での目視による調査を含め、史料の概要を調査する（詳細は第３項に示す。）。

（２）ＡＩ-ＯＣＲ用学習データ作成

契約相手側は、貸与された画像から様々な書式パターンの適切な画像を抽出し、手書き文字計 50 万文字以上の AI(Artificial Intelligence)-OCR(Optical Character Recognition) 学習用データセットを作成する。（詳細は第４項に示す。）。

（３）ＡＩ-ＯＣＲ精度向上実証調査

前項で作成した学習データをＡＩ-ＯＣＲシステムに学習させ、当該システムを開発し、ＯＣＲの精度向上効果について報告書としてまとめる。(詳細は第５項に示す。)

（４）活字識別精度向上策

ＮＤＬ-ＯＣＲを使用した際の活字識別精度向上に資する方策を講じ、この際、追加学習データを作成する場合は、２．１（２）項に含めて提出する。（詳細は第６項に示す。）

（５）ＡＩ-ＯＣＲによるテキスト化

約２，０００点の手書き文字資料及び活字史料に対して、ＡＩ-ＯＣＲを用いてテキスト化処理を実施する。(詳細は第７項に示す。)

（６）生成ＡＩを活用した検索・分析システムの実証環境の提供

契約相手方は、官側がテキスト化された史料の検索や内容の整理・分析を可能にする、生成ＡＩを活用したシステムを利活用できることを確認するための環境を準備する。（詳細は第８項に示す。）

２．２　作業期間

契約締結日から令和８年３月２３日（月）まで

２．３　作業場所

契約相手方作業所等

ただし、官側から要請があった場合は速やかに協議する必要があるため、作業場所は関東甲信越地方に限定する。

２．４　準　拠

契約相手方は、契約後速やかに官側と作業打合せを行い、作業内容等の詳細について確認決定し、作業スケジュール表を作成し、官側の承認を得て作業を実施する。

作業スケジュール表は、作業項目ごとに作業人員、想定処理量、作業期間等を記述した週単位のスケジュールを記載し、速やかに官側に提出すること。

２．５　権　利

本作業で作成された基礎史料調査の報告書、学習用データ、精度向上実証の報告書、活字史料のＯＣＲテキストの編集著作権は官側に帰属する。また学習用データは官側が許諾した場合のみ、ＡＩ－ＯＣＲシステムに学習させることができるものとする。

２．６　秘密保持

本作業を実施するにあたり知り得た情報及び作成したデータ、報告書等の内容については、第三者に洩してはならない。

２．７　個人情報保護

本作業を実施するにあたり、官側が貸与する画像データ内に個人情報を確認した場合、個人情報保護法や関係法令を遵守し、個人情報データの漏洩防止のために必要な安全管理措置を講じつつ、当該史料は官側に返却するものとする。

２．８　成果物

契約相手方は以下に示す成果物を官側に納品するものとする。格納媒体は契約相手方で準備するものとする。細部については、官側との協議に基づき、実施するものとする。

（１）基礎史料調査報告書：紙及びデータ格納媒体［ＤＶＤ-Ｒ］
（２）学習用データ※　　：データ格納媒体　　［ＤＶＤ-Ｒ］※
（３）精度向上実証調査報告書：データ格納媒体［ＤＶＤ-Ｒ］
（４）ＡＩ－ＯＣＲテキスト：データ格納媒体　　［ＤＶＤ-Ｒ］
（５）生成ＡＩを活用した検索・分析システムの利用手順書

※２．１（４）により生じたデータを含む。

２．９　品質保証

（１）官側へ納品した成果物に不具合が発見された場合は、納入後３年間、契約相手方の責任において速やかに修正等の対応を行うものとする。
（２）修正作業等により電子画像データの引渡しを受ける場合は官側の許可を得るものとする。

３　基礎史料調査

３．１　史料現地調査

官側との協議に基づき、官側が保有する史料全点（約７５，０００点、約２，５００万枚）について調査を行い、必要に応じて当研究所Ｆ１棟史料庫において現

地調査を行う。史料の規模、性質、レイアウト、画像の画質、筆跡の難易度などについて調査を行い、ＡＩ－ＯＣＲテキスト化作業を行うにあたって、テキスト化、史料の整理、分類等を含め必要な作業工程や時程などについて報告書にまとめる。またＡＩ－ＯＣＲテキスト化の作業を行う上で効率的なグループに仕分ける

３．２　手書き史料・活字史料の仕分け

官側が保有する史料の画像データにつき、手書き史料と活字史料を自動的に仕分けるシステムまたはプログラムについて検討・製作する。そのシステムまたはプログラムにより、史料約２，０００点(１００万枚)について仕分けを行う。対象画像は３．１項の調査に基づき官側と協議の上、選定するものとする。その際、個人情報が含まれる画像は選外とする。

４　ＡＩ-ＯＣＲ学習データ作成

４．１　基礎史料調査において自動振り分けされ、分類された手書き史料のうち、官側との協議によって先行的に学習データを作成するものについて、崩し字（行書及び草書体)を概ね６割以上含むＡＩ-ＯＣＲ学習データを５０万文字分以上作成する。その際、個人情報が含まれる画像は選外とする。

４．２　ＡＩ-ＯＣＲ学習データセットの仕様を以下に示す。

（１）仕様に未記載の事項で疑問が生じた場合は、官側と協議の上決定するものとする。

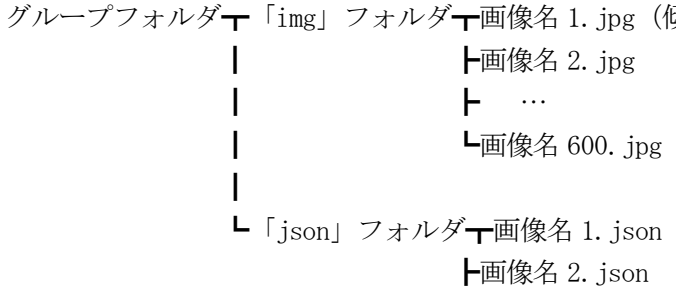
（２）評価用データは、画像データと画像データに対応する JSON データの組み合わせで作成する。

（３）画像データが 0.5° 以上傾いている場合は傾きを補正し、文字が極力正立した状態に修正する。画像データの傾き補正情報は下記の構造で作成し、TSV 形式テキスト（または Excel データ）で納品する。

[TSV 形式テキスト出力例]
画像名 1.jpg[タブ] 0.6[改行]
画像名 2.jpg[タブ]-1.2[改行]
画像名 3.jpg[タブ]-1.2[改行]
…
画像名 600.jpg[タブ]2.0[改行]

（４）JSON データには画像に対応したアノテーション情報(座標情報、テキスト等)を入力する。

（５）学習用データは画像のレイアウト種別等のグループごとにそれぞれ 1 フォルダ（以下：グループフォルダ）に格納し、以下の構造で作成する。



┣ …
┗画像名 600.json

画像補正データ.xlsx

（6）JSON データのテキストは官側が別途指示する翻刻方針を基準として入力する。翻刻方針の変更が必要な場合は、官側と協議の上決定するものとする。

（7）JSON データのテキストは入力精度 99%以上とする。
ただし、史料の破損や汚れ等で判読不能な文字、及び翻刻方針上入力不能な文字については、官側と協議の上、精度保証の対象外とする。

（8）JSON データは、a「資料情報（BOOK)」、b「画像情報（PAGE)」、c「行矩形情報（LINE)」の 3 階層で構成する。

a BOOK の定義
- BOOK は「資料名（TITLE)」の属性を持つ。
- BOOK は子要素として PAGE を持つ。

b PAGE の定義
- PAGE は「対象画像ファイル名（NAME)」「高さ（H)」「横幅（W)」の各属性を持つ。
- PAGE は子要素として LINE を持つ。

c LINE の定義
- LINE は内包する文字の集合の外接矩形とする。
- LINE は「行種別（TYPE)」「書字方向（ALLIGN)」「行内テキスト（STR)」「4 点座標（POINTS)」の各属性を持つ。
- TYPE は、"本文"・"ルビ"のいずれかを入力する。
- ALLIGN は、"縦組"・"横組（左から右）"・"横組（右から左）"・"反転"のいずれかを入力する。
- STR は、行内の文字を OCR または目視で解読したテキストを入力する。テキストは校正することで精度 99%以上を保証する。
- STR に史料の破損や汚れ等で判読不能な文字、及び翻刻方針上入力不能な文字が含まれる場合には当該文字を「=」で入力する。
- POINTS は「x1 y1 x2 y2 x3 y3 x4 y4」の形式で入力する。

[JSON 構造例]

```
BOOK {資料名, [
 PAGE  {幅, 高さ, 画像ファイル名, [
        LINE {座標情報, テキスト, 付加情報※},
        LINE {座標情報, テキスト, 付加情報},
        LINE {座標情報, テキスト, 付加情報}
 ]}
]}
```

※本文/ルビの区分、書字方向情報（縦組、横組（左から右)、横組（右から左)、反転)

[JSON 形式 OCR テキスト出力例]

```
{"BOOK":
   {"TITLE":"海軍省-公文備考-S1-43-3396", "PAGE":[
      {"W":"3348", "H":"4708", "NAME":"KS31_0013_01.TIF", "LINE":[
```

{"POINTS":[20, 340, 940, 340, 940, 1200, 20, 1200], "ALLIGN":"縦組", "TYPE":"ルビ", "STR":"デンキキョウカイ"},
{"POINTS":[2139, 361, 2139, 551, 2269, 551, 2269, 361], "ALLIGN":"縦組", "TYPE":"本文", "STR":"社團法人電氣協會主催ノ下ニ"},
{"POINTS":[1380, 372, 1380, 714, 1540, 714, 1540, 372], "ALLIGN":"縦組", "TYPE":"本文", "STR":"制轉機ノ轉把ヲ廻ハシテ…緩ム"}
]}
]}
}

５　ＡＩ-ＯＣＲ精度向上実証調査

第４項で作成したＡＩ-ＯＣＲ学習データをＡＩ-ＯＣＲに学習させ、学習後のＡＩ-ＯＣＲの処理結果の精度をＦ値（※）及び編集距離で評価する。評価用データについては官側が提供するものとする。使用するＡＩ－ＯＣＲプログラム及び実行環境は契約相手方で準備する。また、ＡＩ－ＯＣＲプログラムは近代の草書を含む手書き文字に対して１００万文字以上学習済みであり、第４項で作成した学習データを追加で学習させた場合の精度向上に関して報告するものとする。

※　国立国会図書館によるＦ値（Fmeasure）の定義は以下の通り。

ytrue　　={正解文字情報に含まれる文字の多重集合},

Ypred　　={認識結果に含まれる文字の多重集合}

Precision=|ypred∩ytrue||ypred|,Recall=|ypred∩ytrue||ytrue|

Fmeasure =2Recall*PrecisionRecall+Precision

また、精度以外の定性的な情報（レイアウト認識の影響等）についても分析する。

評価・分析結果は AI-OCR 精度向上実証調査報告書にまとめて納品する。

６　ＮＤＬ-ＯＣＲプログラムの改修

官側が保有する活字史料のうち、主要なレイアウトを抽出し、レイアウト認識の向上等を目的としたＮＤＬ-ＯＣＲプログラムの改修を行う。レイアウトの抽出のための史料の点数及び種類については官側と協議するものとする。プログラムの改修を行った後、官側が提供する評価データに基づき精度評価を行い、その結果を官側に報告する。その内容については、「ＡＩ-ＯＣＲ精度向上実証調査報告書」に含むものとする。その際、認識精度については中央値 90 パーセント以上を目標とするが、その結果については官側と協議するものとする。また改修にあたって、追加学習データを作成する場合、第４項作業の成果物と併せこれを提出する。

７　ＡＩ－ＯＣＲによるテキスト化

第３項の自動仕分けによって抽出された手書き文字史料及び活字史料に対して、それぞれ８万枚以上の画像を選定し、ＡＩ－ＯＣＲプログラムによるテキスト化を行う。その際、手書き文字史料に対しては第５項で検証したＡＩ－ＯＣＲを使用し、活字史料に対しては第６項で改修したＮＤＬ－ＯＣＲを使用する。

８　生成ＡＩを活用した検索・分析システムの実証環境の提供

クラウド上でセキュリティを担保した生成ＡＩを活用し、第７項のテキストデータ及

び官側が別途支給したテキストデータによる、全文検索等利活用の検証を行うための環境を構築、これの利用手順書を作成し提供する。生成ＡＩ及びそれを利用するサービスについては、官公庁及び地方自治体への導入実績のあるサービスを使用するものとする。

９　成果物、提出書類の納期等

９．１　成果物

成果物引き渡しは表１のとおり。

表１

名　称	数 量	引渡時期	引渡場所
基礎史料調査報告書	１　式	作成完了次第	防衛省 防衛研究所 東京都新宿区市谷本村町５－１
ＡＩ-ＯＣＲ学習データ	１　式		
ＡＩ-ＯＣＲ精度向上実証調査報告書	１　式		
ＡＩ－ＯＣＲテキストデータ	１　式		
生成ＡＩを活用した全文検索システムの利用手順書	１　式		

９．２　その他提出書類

提出書類については表２のとおりとする。

表２

名　　称	部数	提出時期	様式等	提出場所	備　　考
従業員名簿及び資格証明書	１部	契約後速やかに	様式随意及び写し	防衛省防衛研究所 東京都新宿区市谷本村町５－１	本役務に携わる作業員名簿及び作業実績
作業スケジュール表	１部		様式随意		本役務の作業予定

１０　入札資格要件

入札参加希望者の資格要件は以下のとおりとし、要件を満たすことを証明する資料を官側の指定する日までに書面で提出し、承認を得るものとする。

１０．１　実　績

公文書館または公文書館等に類する機関が保有する古文書や公文書の電子化・テキスト化業務に関し、過去５年間でＡＩ-ＯＣＲの学習・評価用データとして活字１０００万字程度又は手書き１００万字程度の役務計３回以上の作成実績及び当該機関向けのＡＩ-ＯＣＲの開発・改修実績を有すること。

１０．２　精度評価レポート

官側が提供する史料２点（手書き史料１点、活字史料１点）に付き、国会図書館が公開しているＮＤＬ－ＯＣＲ及びＮＤＬ－古典籍ＯＣＲを使用し精度評価を行い、その結果を報告書として提出すること。

１０．３　学習データ作成

官側が提供する手書き史料１点に付き、今回の契約において使用する予定のＡＩ－ＯＣＲを使用して学習データを作成し、当該ＡＩ－ＯＣＲの精

度評価の結果と合わせて提出すること。

１０．４　セキュリティ

官側が貸し出す画像の保管場所については入退出管理及び２４時間警備体制が確立され、使用パソコンについては、社外への流出防止処置が施されているものとする。

１０．５　作業従事者

精度評価の報告書作成は、実務経験５年以上の正社員が行うこと。また、当該者は精度評価データ及び学習データ作成の知識を有している、（社）日本文書情報マネジメント協会が行う上級文書情報管理士の資格保有者であること。

現場責任者及び現場副責任者は、上記実務において、責任者又は副責任者等を務めた実績を有する者であること。

１１　その他

この仕様書に疑義が生じた場合は、速やかに官側と協議するものとする。