仕様書			
件　名	戦史史料のＯＣＲ評価用データ等の作成	部　課　名	防衛研究所 戦史研究センター

１．　総　則

１．１　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評価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

官側は、同所が保持する旧陸海軍文書の画像データ（ＣＤ-Ｒ）を作業側に支給する。（詳細は第３項に示す。）契約相手方は、支給された画像データから評価用データに適する画像を抽出、以下の作業を実施する。

（１）翻刻方針策定

支給された史料画像を確認し、対象史料に最適なテキスト作成の方針を策定する。（詳細は第４項に示す。）

（２）ＡＩ-ＯＣＲ評価用データセット作成

作業側は、支給された画像から適切な画像を抽出し、計 3,000 点以上の AI（Artificial Intelligence: AI）-OCR（Optical Character Recognition: OCR）評価用データセットを作成する。（詳細は第５項に示す。）なお、作業の進捗に伴い、翻刻方針に変更の必要が生じた場合は官側と協議のうえ、承認を得た翻刻方針に従い作業を行うものとする。

（３）ＡＩ-ＯＣＲ精度評価

（２）で作成した評価用データを使用し、NDLOCR または NDL 古典籍 OCR に関する精度評価を行い報告する。（詳細は第６項に示す。）

２．２　作業期間

契約締結日から令和７年３月２１日（金）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精度評価データ、翻刻の基準に関する報告書等の編集著作権は官側に帰属する。

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、官側が支給する画像データに個人情報が含まれる場合は、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）翻刻方針レポート：紙及びデータ格納媒体［ＤＶD-R］
（２）評価用データ　　：データ格納媒体［ＤＶD-R］
（３）精度評価レポート：紙及びデータ格納媒体［ＤＶD-R］
※（１）（３）は同一の媒体で提出

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　官の支給する画像データ

官側は下記に示す旧陸海軍文書の画像データ（ＣD-R）を契約相手側に支給する。

画像データの内訳は以下の通りとする。

史料作成時期	明治	大正	昭和
陸　軍	９枚	３枚	６枚
海　軍	９枚	３枚	６枚

当該ＣＤ－Ｒは部外持出しを禁ずる。また、作業終了後は官側に返還する。作業過程で同ＣＤ－Ｒから複製したデータは、官側の指示により消去する。

４　翻刻方針策定

対象史料に最適なテキスト作成の方針（翻刻方針）を策定する。翻刻方針は、史料のレイアウトパターンや漢字の包摂基準、記号・変体仮名・他言語文字の取り扱い方針等を官側と協議の上、凡例化する。

策定した翻刻方針は翻刻方針レポートにまとめて納品する。

５　評価用データセット作成

ＡＩ-ＯＣＲの精度評価・分析に使用する評価用データセットを作成する。

５．１　ＡＩ-ＯＣＲ評価用データセット作成納品データの内訳は以下の通りとする。

作業側は官側が支給する画像データから、評価用データに適する画像を抽出し、評価用データを作成、下記に示す数量以上のデータを納品する。

（1）活字史料・罫線無しレイアウト　　：納品データ：600 点以上
（2）手書き史料・罫線無しレイアウト　　：納品データ：600 点以上
（3）手書き史料・縦罫線ありレイアウト　：納品データ：600 点以上
（4）手書き史料・セル型レイアウト　　：納品データ：600 点以上
（5）活字手書き混在史料・セル型レイアウト：納品データ：600 点以上

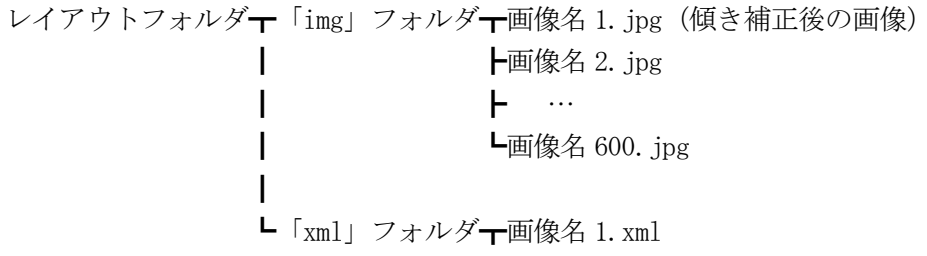
※　画像資料において、1 頁を 1 点、見開き 2 頁は 2 点と換算する。折込の表等は 1 枚を 1 点と換算する。手書き史料には、一定件数以上の草書体（崩し字）画像を含めるものとする。

５．２　ＡＩ-ＯＣＲ評価用データセットの仕様を以下に示す。

（1）仕様に記載されていない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 XML 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）XML 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）評価用データは５．１（1）～（5）項の各レイアウトについて、それぞれ 1 フォルダ（以下：レイアウトフォルダ）に格納し、以下の構造で作成する。



├画像名 2.xml
├　…
└画像名 600.xml

画像補正データ.xlsx

（6）XML データのテキストは第４項の翻刻方針に従い入力する。
（7）XML 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）XML データは、a「画像情報（PAGE)」、b「文字または図版の領域情報（BLOCK)」、c「行矩形情報（LINE)」の 3 階層で構成する。
　a　PAGE の定義
　　- PAGE は「対象画像ファイル名（IMAGENAME)」「高さ（HEIGHT)」「横幅（WIDTH)」の各属性を持つ。
　　- PAGE は子要素として BLOCK を持つ。
　b　BLOCK の定義
　　- BLOCK は書字方向が同一かつ隣接している行矩形情報（LINE）の集合の外接矩形、または図版の外接矩形とする。
　　- BLOCK は「ブロック種別（TYPE)」「4 点座標（POINT)」の属性を持つ。
　　- TYPE には"活字"・"手書き"・"混在"・"図版"のいずれかを入力する。
　　- TYPE が「図版」以外の場合、BLOCK は子要素として「行矩形情報（LINE)」を持つ。
　　- POINT は「x1 y1 x2 y2 x3 y3 x4 y4」の形式で入力する。
　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（DIRECTION)」「行内テキスト（STRING)」「4 点座標（POINT)」の各属性を持つ。
　　- TYPE は、"本文"・"ルビ"のいずれかを入力する。
　　- DIRECTION は、"縦組"・"横組（左から右）"・"横組（右から左）"のいずれかを入力する。
　　- STRING 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。
　　- STRING 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。
　　- POINT は「x1 y1 x2 y2 x3 y3 x4 y4」の形式で入力する。

[XML 構造例]

```
<?xml version="1.0" encoding="utf8" standalone="yes"?>
<OCRDATASET xmlns="NIDS">
<PAGE IMAGENAME="画像名 1.jpg" WIDTH="2000" HEIGHT="3000">
  <BLOCK TYPE="図版" POINTS= "20 340 940 340 940 1200 20 1200"/>
  <BLOCK TYPE="活字|手書き|混在" POINTS= "20 340 940 340 940 1200 20 1200"/>
    <LINE TYPE="本文|ルビ" DIRECTION="縦|横(左から右)|横（右から左）" STRING="テキスト 1" POINTS= "20 340 940 340 940 1200 20 1200"/>
    <LINE TYPE="本文|ルビ" DIRECTION="縦|横(左から右)|横（右から左）" STRING="テキスト 2" POINTS= "20 340 940 340 940 1200 20 1200"/>
```

```
  </BLOCK>
  <BLOCK TYPE="活字|手書き|混在" POINTS= "20 340 940 340 940 1200 20 1200"/>
    <LINE TYPE="本文|ルビ" DIRECTION="縦|横(左から右)|横（右から左）" STRING="テキスト3" POINTS= "20 340 940 340 940 1200 20 1200"/>
    <LINE TYPE="本文|ルビ" DIRECTION="縦|横(左から右)|横（右から左）" STRING="テキスト4" POINTS= "20 340 940 340 940 1200 20 1200"/>
  </BLOCK>
</PAGE>
</OCRDATASET>
```

6　ＡＩ-ＯＣＲ精度評価

作成したＡＩ－ＯＣＲ評価用データ全点をNDLOCRまたはNDL古典籍OCRでOCR処理し、処理結果の精度をF値（※）および編集距離で評価する。

※　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。

NDLOCRとNDL古典籍OCRの実行環境は契約相手方が構築するものとする。

評価・分析結果は精度評価レポートにまとめて納品する。

7　提出物の納期等

（１）作成物引き渡しは表１のとおり。

表１

名　称	数 量	引渡時期	引渡場所
翻刻方針レポート	１　式	作成完了次第 （２０２４年度内）	防衛省 防衛研究所
評価データ	１　式		
ＡＩ-ＯＣＲ　精度評価レポート	１　式		

（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
作業計画表	１部		様式随意		本役務の作業予定

8　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。

（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手書きの古文書や公文書の電子化・テキスト化契約実績、過去３年間でＡＩ-ＯＣＲの学習評価用データとして５０００万字以上の作成実績経験および当該機関向けのＡＩ-ＯＣＲの開発・改修実績を有すること。

（2）精度評価レポート

官側が提供する史料２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。

（3）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。

（4）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およ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。

9　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。