要　旨

本稿は、利用者の資料検索支援を目的として、国立公文書館（以下「館」という。）におけるAI-OCRを用いた資料の全文検索の導入を考察するものである。館では、特定歴史公文書等の利用促進に努めており、利用者へのオンラインサービスである「国立公文書館デジタルアーカイブ」において、所蔵する特定歴史公文書等の目録の検索や資料画像等の閲覧を提供している。しかし、資料内容には目録に記載のない情報も含むため、対象の資料を特定するための言葉が目録上にない場合、中身を閲覧せずに内容を把握することが困難である。また、画像データを提供している場合でも、膨大な資料画像から求める情報を探し出すために端から確認することが必要である。課題への対応として、一つは検索の手がかりとなる有効なキーワードを検索の対象として効率的に増やすこと、もう一つは大量の検索対象から利用者が探している情報を効率的に探し出す方法を提供することが必要である。これらを解決するには、図書館等で大規模な導入例のある、本文のテキスト化による全文検索サービスを提供することが有効と考え、その導入方法として近年発展が目覚ましいAI-OCRの技術に着目した。公文書館の所蔵資料の特徴及びAI-OCR技術の特性の分析を通じて解決策と現行の課題を明らかにするため、諸外国の国立公文書館の先行事例を調査した。

事例より、課題解決の可能性として、資料内容から検索に有効なキーワードと考えられる部分（人名、地名等）にテキスト化範囲を限定してAI-OCRによる効率的なテキスト化を行うこと、そして既存の目録情報、または作成したテキストデータから抽出した固有表現の活用により、有効なキーワードを効率的に増やし、当該キーワードを用いて大量の検索対象から効率的な検索サービスが提供されていることを確認した。また、自動生成したテキストデータに対して、完全一致に限らない検索及び検索結果の提示方法を確認した。AI-OCRによるテキスト化の対象範囲及び検索サービスの提供方法に関する調査結果をもとに、館の所蔵資料について、書式の特徴や活用可能なデータセットの用意、既に提供している目録情報等の現状から対象資料の候補の検討を試みた。

今後の課題として、キーワードの選定や個々の検索オプション、検索結果の示し方等、AI-OCRを用いて作成したテキストデータの精度を考慮した検索手段の提供方法の調査が必要であることを挙げ、全文検索サービスにより発見した断片的な情報を説明するために、既存の目録情報と結びついた検索支援が必要であることを述べた。