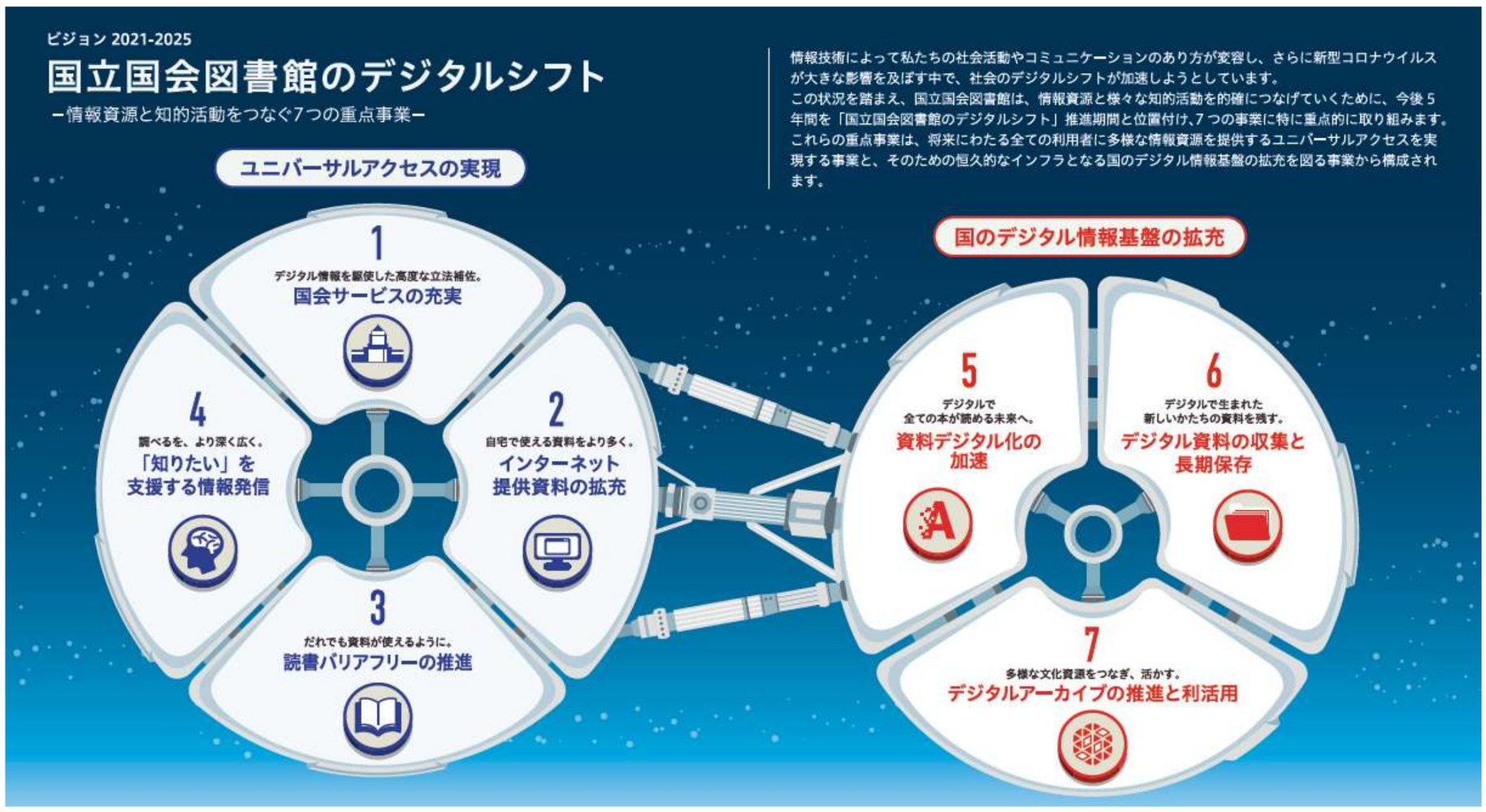
国立国会図書館が保有するデジタル化資料　247万点・2億2300万枚超の全文テキストデータ化に「CLOVA OCR」が採用

2021.07.15　AI関連サービス

『ビジョン2021-2025 国立国会図書館のデジタルシフト』の一環として、
視覚障がい者や高齢者を含む全ユーザーの利便性向上、アクセスの飛躍的な拡大を目指す

LINE株式会社 AIカンパニー（本社：東京都新宿区、カンパニーCEO：砂金 信一郎）は、国立国会図書館（東京都千代田区）が保有する247万点、2億2300万枚を超えるデジタル化資料のOCRテキストデータ化プロジェクトに、CLOVA OCRが採用されましたことを、お知らせいたします。

国立国会図書館では『ビジョン2021-2025 国立国会図書館のデジタルシフト』の一環として、デジタルで全ての国内出版物が読める未来を目指し、来年3月までに247万点のデジタル化資料をテキストデータ化する取り組みが行われています。
『ビジョン2021-2025 国立国会図書館のデジタルシフト』詳細：https://vision2021.ndl.go.jp/

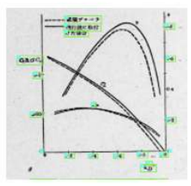


国立国会図書館のデジタル化資料を検索によって簡単に探し出して利用できるようにすること、ひいては、視覚障がい者や高齢者などを含むユーザーにバリアフリーかつ利便性の高い環境を提供し、デジタル化資料へのアクセスを飛躍的に広げることを目指したOCRテキストデータ化プロジェクトに、CLOVA OCRが用いられることになりました。

今回テキストデータ化するデジタル化資料の多くは昭和前期以前の資料であり、レイアウトも複雑なため、学習機能のない既存のOCRでは同プロジェクトに必要な精度に達しないことや、2億2300万枚を超えるデジタル化資料の処理に時間を要する点が課題でした。

CLOVA OCRは文書解析と認識に関する国際会議ICDARで4分野にて世界No.1※を獲得するなど、OCRモデル研究開発に関して経験豊富なチームが開発を行っています。
同プロジェクトで要求される項目に最適なOCRモデル（ルビ、割注、割書きといった特殊な文書に関しても人手を介さず読み取りする、等）を、スピーディーかつ高いクオリティで開発・実現することがCLOVA OCRでは可能です。
※2019/3/29時点

■資料の特性と特化型モデル開発方針（一例）

資料の特性	資料の例	開発方針
認識対象外となる 文字が多い (図及び写真内の文字)		デジタル化資料を利用し、要求事項に沿ったアノテーションデータを作成して学習させることで、認識対象外の文字列が検出される確率を下げるようにします。
ルビ、割注・割書きが多い		テキストボックスの相対的な大きさの差を利用して、ルビ・割注・割書きは別のテキストボックスとして認識するようにします。統計的な手法での後処理により、ルビ・割注・割書きを本文から分離して対応します。
ノイズが多く含まれている 文字領域の検出が 難しい		OCR処理対象となるデジタル化資料をアノテーションしてモデルを学習させることで、認識精度をあげる想定です。

LINEのAIカンパニーでは、AI技術やサービスを通して、生活やビジネスに潜む煩わしさを解消すること、社会機能や生活の質を向上させることで、より便利で豊かな世界をもたらしたいと考えています。「ひとにやさしいAI」が自然なカタチで生活やビジネスの一部となるような、「これからのあたりまえ」を創出するべく、引き続きAI技術のさらなる向上や、ビジネスの連携を進めてまいります。

＜国立国会図書館様コメント＞
今回の事業により当館が入手するテキストデータは、「全文検索」という資料の発見を助けるための検索が主目的ですが、大規模データセットとしてのAI領域での活用や、視覚障害者等の方々の読み上げ利用への期待も高まっています。御社のこれまでの経験を活かして当館のデジタル化資料に最適化させたCLOVA OCRのテキスト化精度に期待しています。

国立国会図書館
National Diet Library, Japan

【国立国会図書館】
国立国会図書館は、「真理がわれらを自由にする」という設立理念に基づき、 国会に属する日本で唯一の国立の図書館として、昭和23年（1948年）に設立されました。 国会の活動をサポートするとともに、国内外の資料や情報を収集した上で永く保存し、 それらの情報資源を広く国民に提供しています。
https://www.ndl.go.jp/

【CLOVA OCRについて】
・CLOVA OCRの認識精度は、横書きや縦書きだけでなく、丸く湾曲して書かれた文字や傾いた文字などの悪条件下での読み取り、多言語の認識、専門用語の認識などで高い精度と評価されました。文書解析と認識に関する国際会議ICDARでは4分野にて世界No.1を獲得しました（2019/3/29時点）。
・フォーマットが決まっている書類はもちろん、あらゆるスタイルの書類を正しくテキスト化します。
・複数枚に及ぶ明細情報の認識が可能です。

【LINE CLOVAについて】
社会に技術とサービスを提供するLINEのAIテクノロジーブランドです。LINEが提供する、文字認識、画像認識、動画解析、音声合成、音声認識といったAI技術やサービスを通して、生活やビジネスに潜む煩わしさを解消すること、社会機能や生活の質を向上させることで、より便利で豊かな世界をもたらしたいと考えています。AI技術が、人に寄り添い、人をサポートし、人の負担を減らす。「ひとにやさしいAI」は、自然なカタチで生活やビジネスの一部となるような、「これからのあたりまえ」を創出します。
https://clova.line.me/

※今回テキストデータ化する資料は、国立国会図書館で広く一般に公開されているものであり、機密データ等は含まれておりません。
