【国立国会図書館の次世代システム開発研究室 における研究活動について　～次世代デジタルライブラリーと資料画像レイアウトデータセットの公開を中心に～】

【木下　貴文】（所属：国立国会図書館　電子情報部電子情報企画課　次世代システム開発研究室）

【発表内容】

国立国会図書館の次世代システム開発研究室（次世代室）では、図書館資料の検索や閲覧機能の向上を目的に、機械学習等の技術を図書館サービスに応用する調査・研究を行い、実験的な Web サービスや学習用データセットを公開している。

本発表で紹介する「次世代デジタルライブラリー」（次世デジ）は、デジタル化資料の検索・閲覧を行うことができる実験的 Web サービスである。OCR 処理によって作成した全文テキストを対象としたキーワード検索や、デジタル化資料の中から自動抽出した画像（挿絵、写真、地図等）を対象とした類似画像検索ができることが最大の特徴である。2021 年 3 月時点の収録資料は、「国立国会図書館デジタルコレクション」でインターネット公開されている著作権保護期間が満了した図書資料・古典籍資料の全て約 33 万 6,000 点（うち、本文テキスト検索の対象は産業分野の約 3 万点であり、デジタル化資料の中から自動抽出した画像は約 2,300 万点）である。その他、デジタル化資料の背景変色部分を自動的に白色化する機能や、スマートフォン・タブレット等の縦長ディスプレイに応じた見開きページの自動分割機能、資料閲覧画面でのページめくり方向の自動判定・設定機能等も搭載しており、それらについても紹介する。

また、次世代室では、2019 年 8 月に GitHub のアカウント（ndl-lab）を開設して以降、次世デジのソースコードや学習用データセットなどの研究成果を公開している。本発表では GitHub 上で公開した成果物の紹介も併せて行う。特に、OCR 処理の精度向上を主たる目的として作成した約 2,300 画像に及ぶ資料画像レイアウトデータセット「NDL-DocL」やその活用事例について説明する。

参考

次世代デジタルライブラリー：https://lab.ndl.go.jp/dl/

ndl-lab GitHub：https://github.com/ndl-lab