※2023年1月18日イベントのディスカッション中に取り上げることができなかったご質問について、登壇者からの回答・コメントを紹介します。（当日チャットで回答・コメントした内容を含みます。）

項番	質問	回答・コメント	回答者
1	NDLOCRについて。数式、図形、グラフにも対応できないものでしょうか。	どのように中身を記述するか、標準が定まっていない分野であることもあり、現時点では中身を読むことを棚上げにしています。「数式」に関しては、数式を含む領域を検出することができ、「図形（グラフを含みます）」も同様に図形を含む領域を検出できます。OCRの読み取った数式の中身を例えばMathXML形式で記述する、といった標準が定まっていき、国内外で利用可能なデータセットが整備・公開されてくれば、それを利用させてもらう形で対応することは考えられます。	青池亨（国立国会図書館）
2	NDLOCRは、スマホで撮影したようなものからの処理も可能ですか。	現在はスマホで撮影したものに対してすぐにNDL OCRを適用できる環境（アプリ）はまだ提供されていないかと思います。ただオープンソースのプログラムなので、そのようなアプリが今後開発される可能性があると思います。	中村覚（東京大学史料編纂所）
		現在利用しているOCRモデルは、計算にそれなりのパワーが必要なので、このままのモデルを利用して実現する場合、処理用のサーバに画像を送信してもらう必要があります。サーバの運用維持や個人情報管理の問題もありますので、当館で実施するのは難しいですが、ソースコードを利用して当館外で実現されるかもしれません。もう1つの方法として、精度は落ちてしまいますがより軽量なモデルを別に作成して、スマホ上で処理を行えるようにすることは考えられます。こちらも当館としてはノープランですが、当館外で実現されると意義のある開発かと思います。	青池亨（国立国会図書館）
3	縦書き日本語書籍のOCR化に関心があります。 市販ソフトには、次の2つの弱点があります。 １．図表に横書きの注や凡例があると、本文の縦書きと横書きのレイアウトの区分けに失敗し文章として意味をなさない結果を出力します。 ２．本文内に海外文献名が登場する場合に、「右に倒れた」アルファベットが縦に並ぶという変則的な状態ですが、こういう認識で失敗します。 NDLOCRではこれらの問題は解決されていますか。	１．図表注については、「キャプション」として、文字列認識の前段階のレイアウト認識時点で分けて検出しますので、正しく検出できていれば、本文とは分けてテキスト化することが可能です。凡例についてはNDLOCRの読み取り対象ではないので、こちらも正しく検出できていれば含まれない可能性が高いと考えます。 ２．縦書きの文献において、海外の文献を参考文献とした場合の文字認識のされ方、と理解しました。こちらについては現時点では文字認識の対象外（テキストが出力されない）としていますが、そうした資料から作成したデータセットを追加で学習することで、認識対象とすることは可能と考えます。	青池亨（国立国会図書館）
4	言語モデルのチューニングなどはどのように実施すればよいのでしょうか。もしも、説明されたもの資料等があれば教えてください。（例えば手書き文字にも対応させたい、手元の読み取り対象に最適化させたいなど）	再学習方法について、現時点でご案内が不足しており申し訳ございません。 ごく簡素な学習手順の説明については下記をご覧ください。 レイアウト認識 https://github.com/ndl-lab/ndl_layout/ 文字列認識 https://github.com/ndl-lab/text_recognition 学習に利用するデータセットの形式や実際のデータセットの例については https://github.com/ndl-lab/pdmocrdataset-part2 をご覧ください。	青池亨（国立国会図書館）
5	以下は、正岡子規の漢詩ですが、左側の行間が詰まっている頁について、句ごとに縦読みで認識せず、（多分頭注と認識しているのだと思いますが）、右側に読んでいきます。「今日‥」の次の句として「吾生‥」となり、次が「仮令‥」となります。これは、正しい順番（上から下、「今日‥」→「万民‥」→「厭世‥」）になるようにするには、NDLOCRの設定を調整すれば（本日青池さんが触れられたような）、簡単に調整することはできるのでしょうか。 https://lab.ndl.go.jp/dl/book/978844?keyword=%E5%AD%90%E8%A6%8F%E5%85%A8%E9%9B%86&keyword=%E9%95%B7%E5%A0%A4%E6%9B%B2%E8%99%95%E5%A2%A8%E6%B1%9F%E9%9A%88&page=35	これは現在の読み順整序モデルで対処できていないレイアウト、ということになります。2023年4月以降に、読み順整序についても改修を行ったバージョンを公開しますので、そちらをお試しいただければと思いますが、特に漢詩の読み順を考慮した手法ではないので、同様に難しいかもしれません。	青池亨（国立国会図書館）
6	グーグルのプログラム実行してみました！ カレンダーなような、数字が多いものを読み込んでみました。結果は曜日の漢字は読み取れましたが、数字は=で読み取れていないようでした。 数字の読み取りは対応していないのでしょうか。	数字も対応しているのですが、縦中横（https://ja.wikipedia.org/wiki/%E7%B8%A6%E4%B8%AD%E6%A8%AA）の認識には対応していないので、恐らく縦中横と認識されたものと考えます。	青池亨（国立国会図書館）
7	中村さんが開発されたGoogle Colabを用いたNDLOCRアプリ，さっそく試させていただき，感銘を受けております。 今手持ちのデータ（PDF・縦書き２段組み）で試してみたのですが，必ずしもうまく読み取ってくれないような気がしました（文字は正確に読み取ってくれるのですが，レイアウトを考慮した読み取り結果にはなっていないといことです。具体的には「奇数ページ上段→奇数ページ下段→偶数ページ上段→偶数ページ下段」という順番に読み取ってくれないということです）。このあたり，使い方について何かコツなどはございますでしょうか。	これは現在の読み順整序モデルで対処できていないレイアウト、ということになります。2023年4月以降に、読み順整序についても改修を行ったバージョンを公開しますので、そちらをお試しいただければと思います。	青池亨（国立国会図書館）
8	NDLOCRのモデルやデータからこちらで学習したモデルを利用した際、本文中の注釈番号((注1) や (*)のようなインデックス)を縦書きの場合ルビと誤認するケースがよくみられます。 上記のようなインデックスをなんと呼ぶべきかわかっていないのですが、本来はどのクラスに分類されるべきで、その点の問題については対策などあったりしますでしょうか。	現在公開しているNDLOCRは、大きな目的として本文に対する全文検索を実現するために開発したものであり、また今年度の追加開発は読み上げへの対応を主目的としているため、読み上げ対象でない要素（もっと言えば本文検索時や読み上げ時にノイズになってしまう要素）については、未検討な部分が多いのが実情です。したがって、ご質問の注釈番号等は未定義である、というお答えになります。 注の内容と本文の位置の対応関係を把握できることは紙面を構造化する目的において有用ですので、公開したNDLOCRを二次利用・拡張する研究として実現されていくことに期待しています。	青池亨（国立国会図書館）
9	NDLがマンガ作品用のOCRモデルを作成する予定はございますか。	今のところ予定はございません。研究として有意義と思いますので、データセットや技術提供の側面でお手伝いできることがあればと思います。	青池亨（国立国会図書館）
10	箱石さんへ　幕末初期であると銀板写真がまだ使われていた時代ですので左右逆さで、ポジネガ反転の文字もあります。それも読み取りはできていますか。	技術的な部分について、専門外の私にお答えすることはできませんが、それができれば良いですね。	箱石大（東京大学史料編纂所）
		現在は読み取れません。文字の左右反転への対応は、誤認識を誘発する可能性がありますので当面は対応できない可能性が高いです。（レアケースよりは、より多数派の紙面の性能向上に優先的に対応する、という方針とご理解ください。）	青池亨（国立国会図書館）
11	箱石さんへ　だれの文字かの筆跡もわかりませんか。	先程の回答と同じになりますが、技術的な部分について、専門外の私にお答えすることはできません。しかし、話題提供の際にもお話ししましたように、いずれ筆者推定の支援ができるようになると良いですね。	箱石大（東京大学史料編纂所）
		OCRで解決すべき課題というよりは、機械学習を用いた人文情報学の研究テーマなのではないかと思います。興味深いですね。	青池亨（国立国会図書館）
12	私は日ごろからOCRをプレゼンの記録に使っていますが、話の進展についてゆく速度が足りないようです。これはどの程度の速度ですか。	こちらについては、OCRよりも音声認識の方が適しているかもしれません。音声認識についても、whisperといった高性能なモデルが公開されているので、そのような取り組みも参考になりましたら幸いです。	中村覚（東京大学史料編纂所）

項番	質問	回答・コメント	回答者
13	箱石さんのご発表にある「検索結果」が「画像のコマ番号」ではなく「刊本の頁数が表示されるとなおよい」ということについて、利用者としても、レファレンスをする立場でも、同意見です。 こうしたことができることによって、研究が格段に進むという実感は持っており、今後の検討課題として位置づけていただければ大変ありがたいです。	NDLOCRでは、ページ番号の読み取りもできます。しかし、紙面上において多くの場合ページ番号は小さく、古い資料は印刷不鮮明な場合もあって必ずしも正確に読めるわけではないので、検索結果としてヒットしたページの画像を目視で確認いただいた方が確実と考えます。	青池亨（国立国会図書館）
14	今後、この技術を使ってOCR化をしたい場合、自館のデジタルアーカイブのPDF等を作成する際の留意事項があればご教示ください。（解像度は●●dpi以上が望ましいといったようなことなど）	高解像度であるほど精度よく読めると思いますが、処理時間や画像のデータサイズが増えてしまうと思いますので、ある程度実験されてから妥協ラインを決めていただくことになると考えます。	青池亨（国立国会図書館）
15	私もフリーのOCRを使っていますが、カメラの能力で随分読み取り精度が左右されるようです。特に自動焦点がうまくピントが合わないで読んでしまうことがあって困っているのですが、それはどのように対策されていますか。	NDLOCRは当館のデジタルコレクションに搭載されている資料の撮影品質を前提に開発したものですので、ピントの合わない資料について、補正等を行うことはできません。撮影時点での品質に気を付けていただく他ないかと思います。	青池亨（国立国会図書館）
16	アンケート自由回答欄のような手書き文字にはどの程度の対応が可能ですか。結構、殴り書きもあります。	手書き文字のOCRに関しては、今回ご紹介したNDL OCRよりも、青池さんから来週（1月24日）公開されるとご紹介があったくずし字OCRの方が有効かもしれません。私としてもとても楽しみにしております。	中村覚（東京大学史料編纂所）
		中村さんにもコメントいただきましたが、縦書きについては、新たに公開しましたNDL古典籍OCR（https://github.com/ndl-lab/ndlkotenocr_cli）をお試しいただくと良いかもしれません。横書きについてはデータセットを整備すればNDL古典籍OCRと同じ仕組みでできる可能性はありますが、まだ具体的な研究課題とはしていません。	青池亨（国立国会図書館）
17	（1月24日公開の）NDL古典籍OCRについて、2点質問があります。 ①絵と文字が組み合わせったもの（浮世絵や読本）も、テキスト化されるのでしょうか。 ②謡本や浄瑠璃の床本などの場合、専門の記号（音符のようなもの）が含まれていると思うのですが、どのようにテキスト化されるのでしょうか。	①はい、絵の中にある文字列についても、検出してテキスト化するようにしています。ただ、通常の紙面よりも難しい条件になりますので、読み落としが増えるかもしれません。 ②「みんなで翻刻」の翻刻成果を加工して学習に利用しているものなので、専門的な記号については翻刻成果物に含まれていれば対応できる、という温度感だと思います。（おそらく、ほとんど含まれないので現時点では対応できていないと思います。）	青池亨（国立国会図書館）
18	2021年度以前のデジタル化資料は業者のOCRとのことですが、遡及的にNDLOCRでかけ直すことはあるのか気になっています。	遡及的にNDLOCRでかけ直すことは想定しておりません。	青池亨（国立国会図書館）
19	現在OCRを手元で動かさずとも、次世代デジタルライブラリー内の書籍であれば直接サイト内からテキストデータを保存できるようにしてくださっていますが、 将来的に国立国会図書館デジタルコレクションの書籍から、同書籍の次世代デジタルライブラリーへのリンク、誘導や、テキスト保存タブなどを実装する予定などはありますか。	次世代デジタルライブラリーは実験サービスであり、そこで有用と判断された機能を国立国会図書館デジタルコレクションに実装していくこととしています。したがって、国立国会図書館デジタルコレクションから次世代デジタルライブラリーへのリンク等の実装の予定はありません。	青池亨（国立国会図書館）
20	箱石さんへ　県立の博物館長（歴史研究者）から、博物館法の改正でデジタルアーカイブ化が義務化されているが、人的体制も予算の措置もない中で、なかなか大変である。デジタルアーカイブ化によって確かに新しい研究手法や新たな研究分野が生まれることに期待したいが、どんな状況か。と聞かれました。 本日の箱石さんのご発表は、こういった質問に対する一つの答えとして、大変参考になりました。	ご感想をお寄せ頂き、誠に有り難うございました。	箱石大（東京大学史料編纂所）