CA1718

動向レビュー

電子化の現場からみたOCRの動向

1. はじめに

インターネットを通じて、自宅や職場などから閲覧できる本が増えている。あるものは無料で、あるものは有料で閲覧することができる。電子化された本がインターネット上で公開される利点は、いつでも／どこでも読むことができるということだけではない。これまで目当ての本を探そうとすると、タイトル、著者名、分類などを頼りに探すしかなかったが、電子化された本は、その中の文章や内容の一部からでも検索可能となる。つまり、インターネット上に電子化された本が公開されると、本の探し方／使い方が変わる、と言える。この新しい「本の探し方／使い方」を陰で支えているのが「光学式文字読取装置」（Optical Character Reader：OCR）というテクノロジーである。本稿では、本や新聞の電子化に携わる者[1]の視点で、OCRの動向を紹介する。

2. OCRはどのように使われているのか

世界各地で進行中の電子化プロジェクトにおいて、OCRはどのような位置づけとなっているのかを確認してみる。

2.1. インターネットで公開する2種類の方法

インターネット上には、電子化された本が、たくさん公開されている。その公開方法には、大きく分けて、2種類ある。

1. 本のページを画像として公開する方法
   例えば、国立国会図書館の近代デジタルライブラリー[2]、米国国立医学図書館[3]など
2. 本のページに書かれた文字をデータとして公開する方法
   例えば、青空文庫[4]、Project Gutenberg[5]など

どちらの方法で公開されていようと、インターネットに繋がっていれば、いつでも、どこからでも閲覧できる、という利点を持つ。「文字をデータとして公開」している場合には、私たちはその内容を直接検索することができる。世界的に見て、本や新聞を電子化している最近のプロジェクトでは、後者の「文字をデータとして公開」することを目標としている場合が多いと思われる。

2.2. 文字データを作成する2種類の方法

文字データを公開することは利点も多いが、その分、コストもかかる。文字データを作成することは、画像を作成するよりも、はるかに手間と時間がかかるからである。その文字データを作成する方法には、大きく分けて、2種類ある。

1. 人手による入力を行う
2. OCRなどのテクノロジーを使う

「人手による入力」という方法で文字データを作成している例として、日本の「青空文庫」が挙げられる。その他、米国において、Distributed Proofreadersというグループが同様の活動を行っている。過去には、Questia社が、「人手による入力」という方法で、大規模に文字データを作成した。一方、OCRなどのテクノロジーを使う例として、Googleブックス[6]や、Amazon.co.jpの「なか見！検索」[7]などが挙げられる。

一般的に、「人手による入力」という方法を使えば、精度の高いものを作ることができるが、時間がかかる。一方、OCRを使えば、効率的に作業が進められるが、精度の点で劣る。欧米でのプロジェクトなどでは、OCRである程度のデータを作成してから、「人手による入力」で修正をしていく方法[8]を採ることもありうるが、それは欧米各国語のOCRが、一般的に精度が高いために可能となっている。日本語のOCRは、欧米各国語に比べて一般的に精度が劣るため、OCRのデータを修正するくらいなら、最初から「人手による入力」を行ったほうが、時間と費用を節約できる場合がある。

2.3. 新しい動き

近年、OCRと「人手による入力」の組み合わせについて、インターネットの特性をうまく活かした「ソーシャル型」とも呼べるような、新しい動きが生まれている。

Australian Newspapers Digitisation Program（ANDP）[9]というオーストラリアの新聞電子化プロジェクトでは、OCRにかけた生データを公開し、それを閲覧者に修正してもらう、という形をとっている[10]。

また、米国議会図書館や米国国立公文書館と提携して電子化を進めているFootnote社では、ユーザーが画像内の手書き文字などを読み取り、画像内の該当する箇所に、付箋を貼れるようなサービスを提供している[11]。

さらに、“reCAPTCHA”（E662参照）というものがある。詳しくは、サイボウズ・ラボ社の秋元氏のブログに書かれている[12]。非常に複雑に組み込まれたシステムであるが、簡単に説明すると、Webサービスの認証プロセスを利用して、OCRの誤変換を訂正していくシステムである。こちらも一種の「ソーシャル型」と言える。

このように、文字データの作成は、コストと時間がかかるので、各社・各機関が様々な工夫を凝らしてくる領域である。後述するように、OCRは完璧ではない。したがって、OCRの動向という場合、実際の電子化現場において、OCRの限界を補うために、「ソーシャル型」をはじめとして、どのような工夫がされているのか、ということまで目を配っておいたほうが良いであろう。

3. OCRの精度に関連する最近の報告

「OCRの精度」というものについて、よく質問を受けるので、一般的にOCRの精度に対する関心は高いと思われる。現在進行中の電子化プロジェクトから、OCRの精度に関連する考察がいくつか提示されているので、その一部をここで紹介する。以下の論文で論じられているように、そもそも「精度の測定」自体が難しいという事情が存在するが、各社・各機関はOCRの精度を高めるために様々な工夫を実施している。

3.1. OCRの精度とは何か？

タナー（Simon Tanner）の論文[13]は、そもそもOCRの精度とは何であり、どのように測定するのか、ということを論じる（E960参照）。そして、彼らの提案する測定方法を用いて、実際に、英国図書館が電子化した新聞コレクションのOCR結果を測定した。特筆すべき点として、精度を測定する際、OCRのC（Character）が示すような「文字（Characters）単位」での正確さだけでなく、「単語（Words）単位」、「ストップワード（検索対象から外す機能語など）を除いた単語（Significant words）単位」、「固有名詞など、大文字から始まる単語（Significant words with capital letter start）単位」、「数字（Number groups）単位」を含む5種類の正確さを検討していることが挙げられる。そして、人名・地名などの固有名詞が多く含まれるような対象は、OCR精度が低くなる可能性を指摘している。

3.2. OCRの精度を上げるために

クリーン（Edwin Klijn）の論文[14]は、現在オランダで進行中の新聞電子化プロジェクトを開始する前に、マイクロフィルムの電子化、JPEG 2000を含むファイルフォーマット、OCRなどに関して、世界の状況を調べたサーベイ論文である。2008年時点での状況を知ることができる。中でもOCRの精度向上に関して、スキャンをカラーで行うか、それともグレースケール（モノクロ写真のように白と黒の間に、段階的な灰色の階調があるもの）で行うか、ということに関して、業者間で意見の相違が見られるとし、ある業者の話として、カラーでスキャンした方が、より良いOCR結果を得ることができると述べている。ただし、あくまで業者の話であって、実際に比べてみたわけではないので、注意が必要である。

パウエル（Tracy Powell）らは、ニュージーランドで新聞電子化を進めるにあたって、グレースケール画像でのOCR変換の是非を論じている[15]。これまで、ニュージーランドにおける新聞電子化は、2値（通常のファックスのように、白黒だけで表現されたもの）でのスキャンを行っていたが、もし、グレースケールでスキャンをした場合、OCRの精度が向上するのかどうかを、コストとのバランスを見ながら、詳細に検討している。その結果は、グレースケールにしたところで、たいした改善は見られない、というものであった。著者らも注意書きをしているが、この結果は必ずしも、グレースケールよりも2値が良い、ということを意味しない。もしすでに2値でスキャンした画像を持っているなら、わざわざグレースケールで再スキャンをする必要はない、ということを示しているだけである。

ホリー（Rose Holly）は、現在オーストラリアで進行中の新聞電子化プロジェクト（ANDP）に関して、OCRの精度を上げるために、「何ができるか」を検討し、そのいくつかを実際に試してみた結果を報告している[16]。著者らが検討したOCRの精度を向上させるかもしれない13個の方法は、原本の選択から、スキャン方法、ファイルフォーマット、画像処理、OCRソフト選定、OCR処理後の修正などを含んでいて、包括的なリストとして参考になるだろう。著者らは、いろいろな方法を組み合わせれば、より良い結果を得ることができる、と結論している[17]。

4. OCRは使えるか？

現在のところ、OCRによる変換は、間違いを伴う。それゆえ、OCRは使い物にならない、という話をよく聞く。ところが、使い物にならないか、それとも使えるのか、というのは「使い方」による。つまり、OCRで作成されたテキストデータをどのように使うのか、という用途次第である。

話を分かりやすくするために、極端な例を紹介する。

筆者の知る限り、2003年10月に米国Amazon社が、“Search Inside”[18]というサービスを発表するまで、「公開するテキストデータは、正確でないといけない」という通念があった。例えば、Questia社のプロジェクトなどが、その考えに忠実なプロジェクトの例である。

ところが、米国Amazon社は、「内容を検索する目

的であれば、OCR結果は必ずしも正確でなくても良い」という新しい考えを持ち込んだ、と筆者は考える[19]。いろいろな解釈がありうるが、おおよそ以下のような理由を挙げておく。

・それ以前は検索が不可能であったのだから、多少の誤変換が含まれていようが、たった1冊でも検索にひっかかるようになるならば、それは前進である。
・もし検索しているキーワードが重要な単語ならば、探している本の中で、繰り返し現れてくるはずである。多少の誤変換があったとしても、本1冊の中で、どこかの部分がちゃんと変換される可能性がある。ひとつでもちゃんと変換されていれば、検索でその本はヒットする。
・さらに、誤変換の中には、「m（エム）」と「r n（アール　エヌ）」[20]に代表されるような「予想できる誤変換」というのがある。このような頻繁に起こる誤変換は、検索システムが処理をすることで、検索の漏れを減らし、再現率を上げることができるかもしれない。

このように、2003年以降、検索目的のプロジェクトでは、OCR変換したテキストデータを、修正することなしに公開する例が増えている[21]。その意味で、OCRは十分に使えるテクノロジーであると言えるが、これはOCRの性能が十分高くなったからではない。誤変換を含むOCR結果を使いこなす方法が見つかったからである、ということに留意する必要がある。

5. おわりに

電子図書館や電子書籍などが盛り上がりを見せている。これから出版される本などは、間違いなくテキストデータでの公開を伴い、検索可能な状態になるはずである。それと同時に、過去に出版された本などを、いかにテキストデータ化して合流させるのか、ということがますます重要な課題になってくる。本稿では、そのような背景を踏まえて、OCRというテクノロジーが、現状どのように使われていて、その精度はどのように考えられているのか、ということを、電子化の現場からの視点で見てきた。OCRの技術者や、販売業者などは、当然、異なる見解を持っているはずなので、本稿はOCRというテクノロジーの一側面だけを紹介した、ということに留意してほしい。

（http://denshikA.cc：denshikA）

(1) 筆者のプロフィールは、以下を参照。
“自己紹介”. denshikA. http://denshikA.cc/profile.php,（参照 2010-05-14）.
(2) “近代デジタルライブラリー”. 国立国会図書館. http://kindai.ndl.go.jp/,（参照 2010-05-14）.
(3) “Turning The Pages Online”. National Library of Medicine. http://archive.nlm.nih.gov/proj/ttp/books.htm,（accessed 2010-05-14）.
(4) 青空文庫. http://www.aozora.gr.jp/,（参照 2010-05-14）.
(5) Project Gutenberg. http://www.gutenberg.org/,（accessed 2010-05-14）.
(6) Google ブックス. http://books.google.co.jp/,（参照 2010-05-14）.
(7) “なか見！検索”. Amazon.co.jp. http://www.amazon.co.jp/b?node=15749671,（参照 2010-05-14）.
(8) 前出のQuestia社はこの方法を採用した。
(9) Australian Newspapers Digitisation Program. http://www.nla.gov.au/ndp/,（accessed 2010-05-14）.
詳しくは、以下を参照。
denshikA.“全豪新聞電子化プログラム”. 電子化. 2009-08-28. http://d.hatena.ne.jp/denshikA/20090828,（参照 2010-05-14）.
(10) Holley, Rose. “Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers”. National Library of Australia. http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf,（accessed 2010-05-14）.
(11) 詳しくは、以下を参照。
bookscanner.“電子化と怒った歴史家”. bookscanner記. 2007-01-27. http://d.hatena.ne.jp/bookscanner/20070127,（accessed 2010-05-14）.
(12) 秋元. “reCAPTCHA - キャプチャを利用した人力高性能OCR”. 秋元@サイボウズラボ・プログラマー・ブログ. 2007-05-25. http://labs.cybozu.co.jp/blog/akky/archives/2007/05/recaptcha-human-group-ocr.html,（参照 2010-05-14）.
(13) Tanner, Simon et al. Measuring mass text digitization quality and usefulness: Lessons learned from assessing the OCR accuracy of the British Library's 19th century online newspaper archive. D-Lib Magazine. 2009, 15(7/8). http://www.dlib.org/dlib/july09/munoz/07munoz.html,（accessed 2010-05-14）.
(14) Klijn, Edwin. The current state-of-art in newspaper digitization: A market perspective. D-Lib Magazine. 2008, 14(1/2). http://www.dlib.org/dlib/january08/klijn/01klijn.html,（accessed 2010-05-14）.
(15) Powell, Tracy et al. Going grey?: Comparing the OCR accuracy levels of bitonal and greyscale images. D-Lib Magazine. 2009, 15(3/4). http://www.dlib.org/dlib/march09/powell/03powell.html,（accessed 2010-05-14）.
(16) Holley, Rose. How good can it get?: Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine. 2009, 15(3/4). http://www.dlib.org/dlib/march09/holley/03holley.html,（accessed 2010-05-14）.
(17) ホリーは、OCRの精度を上げるために、あれこれと試してみたが、一番効果的なのは、技術的なものではなく、「人手による入力」によって修正をする、という方法であった、ということも述べていて、本稿2.3.で紹介した「OCRにかけた生データを公開し、それを閲覧者に修正してもらう」というANDPの手法の有効性を示していることにも留意する必要がある。
(18) “Search Inside the Book”. Amazon.com. http://www.amazon.com/b?node=10197021,（accessed 2010-05-14）.
(19) bookscanner.“証人喚問前半”. bookscanner記. 2007-01-31. http://d.hatena.ne.jp/bookscanner/20070131,（参照 2010-05-14）.
bookscanner.“証人喚問後半”. bookscanner記. 2007-02-02. http://d.hatena.ne.jp/bookscanner/20070202,（参照 2010-05-14）.
少しふざけた調子で書かれているが、大筋を理解する程度には正確な内容となっている。

(20) 例えば、「make」（エム、エイ、ケイ、イー）は頻繁に「rnake（アール、エヌ、エイ、ケイ、イー）」と誤変換される。「rnake（アール、エヌ、エイ、ケイ、イー）」で、検索してみると、いくつかヒットする。
“rnake”. Google ブックス .
http://books.google.co.jp/books?q=rnake,（参照 2010-05-14）.
(21) 例えば、Google ブックス、Internet Archive など。

CA1719

動向レビュー

デジタルゲームのアーカイブについて ―国際的な動向とその本質的な課題―

1. Before It's Too Late

世界的に活動しているゲーム開発者、研究者のNPO組織「国際ゲーム開発者協会」（International Game Developers Association：IGDA）の専門部会である「ゲーム保存研究会」（Game Preservation SIG）は、2009年3月、最近のデジタルゲーム保存の現状と課題についての白書（以下、「ゲーム保存白書」と称する）を取りまとめた。それは、次のような書き出しで始まっている。

「デジタルゲームの保存は急を要している。毎年、何千ものゲームが、他のすべてのデジタルメディアを脅かしている寿命の問題、すなわち情報の欠落と旧式化によって失われつつある。デジタルメディアは、原材料の経年劣化によって驚くほど寿命が短く、メディアフォーマットが絶えず変化するために急速に陳腐化する。そして、それらを動かすためのハードウェアも同様である。」(1)

デジタルゲームは、マンガやアニメーションとならんで現代のポップカルチャーを代表する表現文化であり、コンテンツであるが、その収集、保存、伝承が極めて危機的な状況にあることは間違いない。その危機は、他の分野が主な保存形式としている紙媒体、画像、映像、文字記録などより、おそらくはるかに深刻な状況である。同白書にも整理されているように、それはデジタルゲームの多くがメディア（ROMカートリッジ、磁気ディスク、光学ディスク等）とその再生装置（ゲームプラットフォーム）の組み合わせによって成立しており、そのそれぞれが物理的、技術的、法的な意味において、現実性のある長期保存のプロセスを確定することを困難にしていることに起因する(2)。

また、同白書は保存する対象を「デジタルゲーム」と表記しているが、情報通信技術の劇的な変革に伴い、デジタル技術を基盤とするゲームとして認識される対象は、いわゆるビデオゲーム（日本ではテレビゲーム）やアーケードゲームだけでなく、オンラインゲームや携帯電話等によるモバイルゲームなどを含む多メディアに展開し、またその内容についても、ごくシンプルなライトゲームから重厚なストーリーと世界観を備えた複雑なインタラクティブ・エンタテインメントまで、非常に大きな振幅を持つよ