ドキュメントスキャン PDF のコストと品質について

既刊本のデジタル化では、ドキュメントスキャンによる「自炊 PDF」がユーザーの支持を受けている。中でもまとまった量の PDF 化では、BOOKSCAN などいわゆる「自炊代行業者」に依頼するケースが増えている。

代行業者はわずか 1〜2 年の間に全国で 110 社を超え、多くの利用者からのニーズがあることを物語っている。（ http://www.bookfire.net/ 調べ）

ここでは代表的な自炊代行業者「BOOKSCAN」によって作成された PDF の品質とファイル形式、価格体系などを分析する。

BOOKSCAN（合同会社大和印刷）
http://www.bookscan.co.jp/
従業員：シフト制の常勤が約 100 名、在宅の目視スタッフが約 100 名。
使用ドキュメントスキャナ：Canon DR-X10C 相当（価格 1,880,000 円）

ファイル形式：スキャン PDF（OCR による透明テキスト付加）
OCR テキストが不要の場合は JPEG や TIFF 画像による納品も可能。

PDF バージョン：PDF1.3（Acrobat 4.0 以上。オープン仕様）
画像圧縮形式：JPEG（高品質）
目次・しおり機能：なし
フォント埋めこみなし

解像度：カラー／グレー：300dpi

自炊用として普及している個人用ドキュメントスキャナ「ScanSnap 1500」の場合、ノーマルが 150dpi、最高のスーパーファインで 300dpi である。また、同じ 300dpi でも画像圧縮の設定値によって画質が大幅に違う。キヤノンやマイクロソフトなどでは、OCR 用として推奨する解像度は 200〜300dpi が効率良いとしている。
http://cweb.canon.jp/e-support/products/canoscan/reading.html
http://office.microsoft.com/ja-jp/help/HA001102160.aspx

参考（ここでは ppi と dpi はほぼ同じものと考える）

一般的な PC 用液晶モニタ	96ppi
iPhone3G	163ppi
iPhone4	326ppi（Retina）
iPad	132ppi
レーザープリンタ	300〜600dpi
商業印刷物	1200〜2400dpi

ファイル容量：元書籍によって違うが、BOOKSCAN によれば、書籍一冊あたり 75MB くらい。辞書などは 3GB くらいにもなる。

・解像度が 2 倍になるとファイルサイズはおよそ 3〜4 倍になる。

・PDF の最終モードが RGB となってしまうため、カラーとグレースケールによるファイルサイズの差はほとんど無い。

価格・コスト：顧客が個人の場合、断裁・スキャン・傾きや抜けの目視チェック・スキャン後の書籍は溶解廃棄。以上で 1 冊 100 円。ただし OCR なしで 350 ページまで。

実際には OCR（100 円）やファイル名を書名に変更するサービス（50 円）などをセットで注文するケースが多いとのこと。

顧客が出版社の場合

■カバースキャンあり（帯状スキャン）1 冊あたり￥900（税抜）

■カバースキャンなし　1 冊あたり￥500（税抜）

共通するオプション

・ページ数カウント

・OCR（透明テキスト）

・名前変更

・納期一週間以内

・全ページ目視チェック

・納品後 10 日以内ならば、再スキャン可能（弊社基準を満たさない場合）

という見積もりが出ている。

［考察］ドキュメントスキャンに求められる品質

コストと納期の他に、スキャン PDF 化に求められるクオリティには次のようなものがある。

・画像品質
　ピントのシャープさ、ノイズの少なさ、拡大率（解像度）
・傾きがないか
・抜け・飛ばし・折れがないか

画像品質や OCR の認識率はシステム的な問題なので、高度な機器（ハード＆ソフト）を使用することで性能を上げることは可能。大量に連続スキャンする場合は、断裁品質（ゆがみや巻き込みが起きないかどうか）や、シートフィーダーによる紙送り圧力の調整などに、業者としての能力差が表れると予想される。

OCR の認識率

PDF といっても、ドキュメントスキャナによるいわゆる自炊 PDF は、InDesign などの印刷用 DTP データから書き出すフォント埋めこみ型の PDF とは別物で、スキャンした画像をページ順にまとめ、場合によって、別レイヤーに OCR による機械読み取りテキストを透明なフォントで乗せるというもの。そのため、OCR にかかりにくい表組みや飾り文字、白抜き文字などは誤認識が増えてしまう。

認識率は元書籍のレイアウトやフォント、カラーなどによって大きく異なるが、ここでは検証用の書籍 11 冊（すべてポット出版刊）を BOOKSCAN に PDF 化させたもので、認識率の照合分析を行った。

校正業者に依頼した今回のチェックの結果、OCR 透明テキストの精度は 97％と高い数字であった。読み上げ用のテキストとしては難しいが、本文の検索用としては使用に耐えるという評価である。

詳細は別紙Ⓐ「OCR 処理生成によるテキストデータと現物書籍との照合分析レポート」を参照。

[参考資料]

ドキュメントスキャン以外の既刊本デジタル化・ケーススタディ

A）　国立国会図書館、所蔵資料の媒体変換

【分量（および期間等の仕様諸元)】
「上掲の候補資料群の中から可能な限り媒体変換を行う」ので、現時点では不明。予定内容は以下の通り。期間は 3 年間（2009～2011）

媒体変換候補資料

ア．保存及び電子図書館サービスの観点からデジタル化を行う資料
　基本的にインターネットで提供することを前提とする。

　帝国議会会議録（約 6 万 2,000 頁）
　明治期、大正期及び昭和前期刊行図書（約 26 万 1,000 冊）
　（※デジタル化完了後に著作権処理を行い、近代デジタルライブラリー及び児童書デジタルライブラリーで追加公開する）
　電子展示会
　古典籍（約 30 万冊）
　国内博士論文（約 50 万タイトル）
　（※「国立国会図書館と大学図書館との連絡会」ワーキング・グループの中間報告に基づき実施する）
　官報[明治 16 年～昭和 22 年分]（約 75 万頁）

イ．保存の観点からデジタル化を行う資料
　当面インターネット提供は行わず、館内のみの提供とする。

　国内刊行和雑誌（約 14 万タイトル）
　（※紙質あるいは頻繁な利用により、資料の劣化損傷状況が著しい、又は、劣化損傷の大幅な進行が予想される資料を対象にデジタル化を行う。）
　昭和 20 年代刊行和図書（約 10 万タイトル）
　（※昭和 20 年代に刊行された官庁出版物及び児童書を対象にデジタル化を行う。紙質が悪い上に閲覧・複写希望が多く、劣化が深刻であるため。）

［参考資料］

ウ．保存の観点からマイクロ化を行う資料

　これまでのマイクロ化の経緯に鑑みて、一定の区切りまでマイクロ化を行うのが適当な資料、及び、外部機関との関係においてマイクロ化が必要とされる資料等についてマイクロ化を行う。

国内刊行新聞（54 タイトル）
（※日本新聞教育文化財団との間の寄託契約による）
旧函架等大型本（約 6,000 コマ）
ＮＤＣＺ（1945 年～1966 年刊行和雑誌）（約 36 万コマ）
他機関所蔵マイクロ新聞（平成 20 年度末現在 115 タイトル）
（※相手機関及び著作権者から、既にマイクロ化に係る内諾又は許諾書の提供を受けているもの）
大正・昭和期刊行寄贈新聞の未整理分及び欠号補充分（約 1 万 5,000 コマ）。

【コメント】
国内、書籍、という条件では最大規模か。また、
全文テキスト化実証実験報告書 | 国立国会図書館 National Diet Library
http://www.ndl.go.jp/jp/aboutus/digitization_fulltextreport.html
において、国内ではほぼ唯一、OCR による大量書籍のテキスト化の調査報告を行っている。

【ポインタ】
・資料デジタル化について | 国立国会図書館-National Diet Library
http://www.ndl.go.jp/jp/aboutus/digitization.html
・プロジェクト全体については、
http://www.ndl.go.jp/jp/aboutus/digitization/zenbun.pdf を参照
・スキャン条件・OCR 作業等については、
http://www.ndl.go.jp/jp/aboutus/digitization/ocrzenbun.pdf を参照

B）近代デジタルライブラリー

【分量（および期間等の仕様諸元）】
・明治期刊行図書　　約 164,000 冊
・大正期刊行図書　　約 96,000 冊
・昭和前期刊行図書　約 310,000 冊

[参考資料]

詳細が公開されていないが、KDI 発表のリリースでは、「平成 12 年度から当館が所蔵する明治期に刊行された図書の全分野を対象として著作権調査を行いました。著作権保護期間が満了したもの、および、著作権者の許諾を得たものからデジタル化に取り組み、 平成 14 年度から近代デジタルライブラリーでの提供を開始しました。」とある。

……ということは、2 年間で 57 万冊（「平成 23 年 2 月現在、47 万冊利用可能」と発表）？

【コメント】
希少本も多いはずなので、「非破壊」でやっているはず。それでこのスピードはすさまじい。
著作権処理が終わっていないものは館外では閲覧できず、というモデルでやっている、ということは、著作権保護期間内の書籍もすでに大量にスキャンしているわけで、著作権処理の部分を出版社がきちんと対応できれば、「出版デジタル機構」は、「近デジ」「国会図書館」にスキャン作業をまかせてしまい（データを使わせてもらい)、「機構」は著作権処理と課金処理だけ行う、という考え方もあるのではないか?

【ポインタ】
このデータベースについて ｜ 近代デジタルライブラリー ｜ 国立国会図書館 -
http://kindai.ndl.go.jp/information/aboutKDL.html

C）CiNii 論文情報ナビゲータ

【分量（および期間等の仕様諸元)】
約 300 万本 ※学術論文なので、本に換算すると数十分の一、くらいか?

【コメント】
2007 年から検索エンジンへの公開をはじめ、アクセスは劇的に伸びたものの一旦停滞し、2009 年にサイトの全面リニューアルとともに web API を公開し、アクセスがさらに倍増している。
情報の公開とともに、書誌情報や書誌構造の提供、といった要素が重要であることを示唆している。
担当者に直接聞いた範囲では、
・業者は、
株式会社ムサシ http://www.musashinet.co.jp/ と
廣済堂 - http://www.kosaido.co.jp/ が担当した模様

・OCR による全文テキスト化は、検索のためと割り切って使っている（スニペットとしてすら利用せず、主に出さないようにしている)。一方で、学術文書であるにもかかわらず、OCR 品質で苦情が来たことは（たぶん）ない。

[参考資料]

ただ、その場にいた専門家からは、単に全文化テキストを持たせただけでは、web のようにリンク／被リンク、タグによる重みを与えたりすることができないので、期待どおりの検索結果にたどりつけないケースが増える、という指摘もなされた（そのためには、目次部分や要約のみを別途切り出すといった構造化／重みづけ処理をすることが有効、とも）。

【ポインタ】
第 5 回：CiNii の挑戦：ボーン・デジタルの情報学｜美術館・アート情報 artscape -
http://artscape.jp/study/born-digital/1213228_2772.html

D）その他、雑多なもの

●Green Apple（ http://www.ebookmatrix.net/ ）
中国の書籍・ドキュメント電子化企業
→ eBookMatrix は、「3 年前（?）」にグリーンアップルを買収した会社らしい（グリーンアップル自体は子会社として存続している）。
・「1989 年創業以来、10 万ページの新聞紙面のデジタル化」
・「（データ入力を行う）300 名の従業員が、99.99%の高精度、かつ米国平均の 1/3 の人件費で働く」
・2001 年からは米国市場に参入　あたりがポイントか。
eBookMatrix.net - The Most Professional Bookstore in Universe -
http://www.ebookmatrix.net/info/info.html

●O-RID KYBER http://www.o-rid.com/jp/company/
→日本で、「人力入力」サービスを展開（おそらく作業は中国内）
名刺の入力料金が 100 件／1200 円、ということは一件 100 文字と仮定して一文字 0.12 円以下、になるから、0.35 円〜の国内相場と比べれば格段に安い。上記の「1/3 の人件費」ともおおよそ符合する。

●Google Books への「入稿仕様」
http://books.google.com/support/partner/bin/answer.py?answer=107073
→「400ppi バイナリ」という表現が意味不明だが、後半の説明からすると、200ppi でも OCR 処理できる、ということらしい

●国内「自炊」業者まとめリンク http://www.bookfire.net/
→ 現在登録業者数 113 件。こまかい条件はあるものの、一冊 100〜200 円がおおむねの相場として確立している。

［参考資料］

●wikipedia における「Book scanning」の定義

Book scanning - Wikipedia, the free encyclopedia -
http://en.wikipedia.org/wiki/Book_scanning
→ 以下のような、基本的な参照先や議論ポイントがまとまっている。

・Project Gutenberg - Wikipedia, the free encyclopedia -
http://en.wikipedia.org/wiki/Project_Gutenberg
・Google Books - Wikipedia, the free encyclopedia -
http://en.wikipedia.org/wiki/Google_Book_Search
・Open Content Alliance - Wikipedia, the free encyclopedia -
http://en.wikipedia.org/wiki/Open_Content_Alliance
・商用「ブック」スキャナ
・大規模書籍スキャニング概論
・破壊（断裁）／非破壊スキャニング

●Scan This Book! - New York Times -

http://www.nytimes.com/2006/05/14/magazine/14publishing.html?pagewanted=all&ei=5090&en=c07443d368771bb8&ex=1305259200
→書籍のデジタル化、について本自体の歴史的な成り立ちから著作権問題（open public でない本をどうするか問題）、Google Books についてまで広く論じている。2006 年の論考。

●Mass Digitization of Books, by Karen Coyle

http://www.kcoyle.net/jal-32-6.html
→ 大量の本をスキャンすること、についての具体例が豊富。Google Books 以外でも、冊数、ペース、業者まで比較的詳しい数字がまとめて載っている。
例：「ミシガン州の図書館で 7 万冊のスキャンは 6 年以内に完了するだろうと述べた」「オペレータが一日約 50 冊をデジタル化可能→ミシガン州の全 700 万冊をデジタル化するのに 20 人がかりで 19 年かかるだろう。」「OCA は、現在月間約 1 万冊をスキャンしている」
「Microsoft が英国図書館の 10 万冊をスキャン」「(非・大量デジタル化の例として）プロジェクト・グーテンベルグは 35 年の間、年間 500 点未満のデジタルテキストを作成」etc. ほか、カーネギーメロン大学、スタンフォード大学、Amazon の取り組み等をコンパクトに紹介。スキャナについては、Kitras Technology と 4DigitalBooks を紹介。

資料作成：ポット SD　日高崇

［別紙 Ⓐ ］

OCR 処理生成によるテキストデータと現物書籍との照合分析レポート（2011/10/18）

【目的】

このレポートは、書籍データをスキャニングして読み込み、OCR 処理を施されて生成されたテキストデータ（以下、OCR テキストデータ）の精度と、OCR 処理時の傾向を調査し明らかにすることを目的とする。

「書籍の種類」（単行本、雑誌、図版、カタログ等)、「ページの性質の種類」（縦組、横組、段組、見出し・脚注等あり、図表あり、目次等)、「文字種類」（記号、日本語数字、英字、ギリシャ文字等）などに場合を分類して OCR 処理における生成精度・傾向を明らかにすることを調査射程とする。

このレポートでは、「全体文字数と誤字数（の比率)」「誤字の発生事例から読み取れる傾向」を特に明確にするものとする。

同時に、ページの性質の種類に関して、照合事例から予測を立てて次の調査を行うための指針とする。

【対象書籍】

『石塚さん、書店営業にきました。』『本の現場』『図書館という軌跡』『出版流通合理化構想の検証』『千代田図書館とは何か』『出版時評 ながおかの意見 1994-2002』『どすこい出版流通』『図書館の近代』『日本の出版流通における書誌情報・物流情報のデジタル化とその歴史的意義』『デジタルコンテンツをめぐる現状報告』『ず・ぼん 16 号』（以上、全てポット出版刊）

【調査方法】

上記 11 冊の書籍の本文 20 ページ分の OCR テキストデータに対して、校正者により現物書籍を元原稿として視認照合による校正を行い、差異を赤字で記入する。

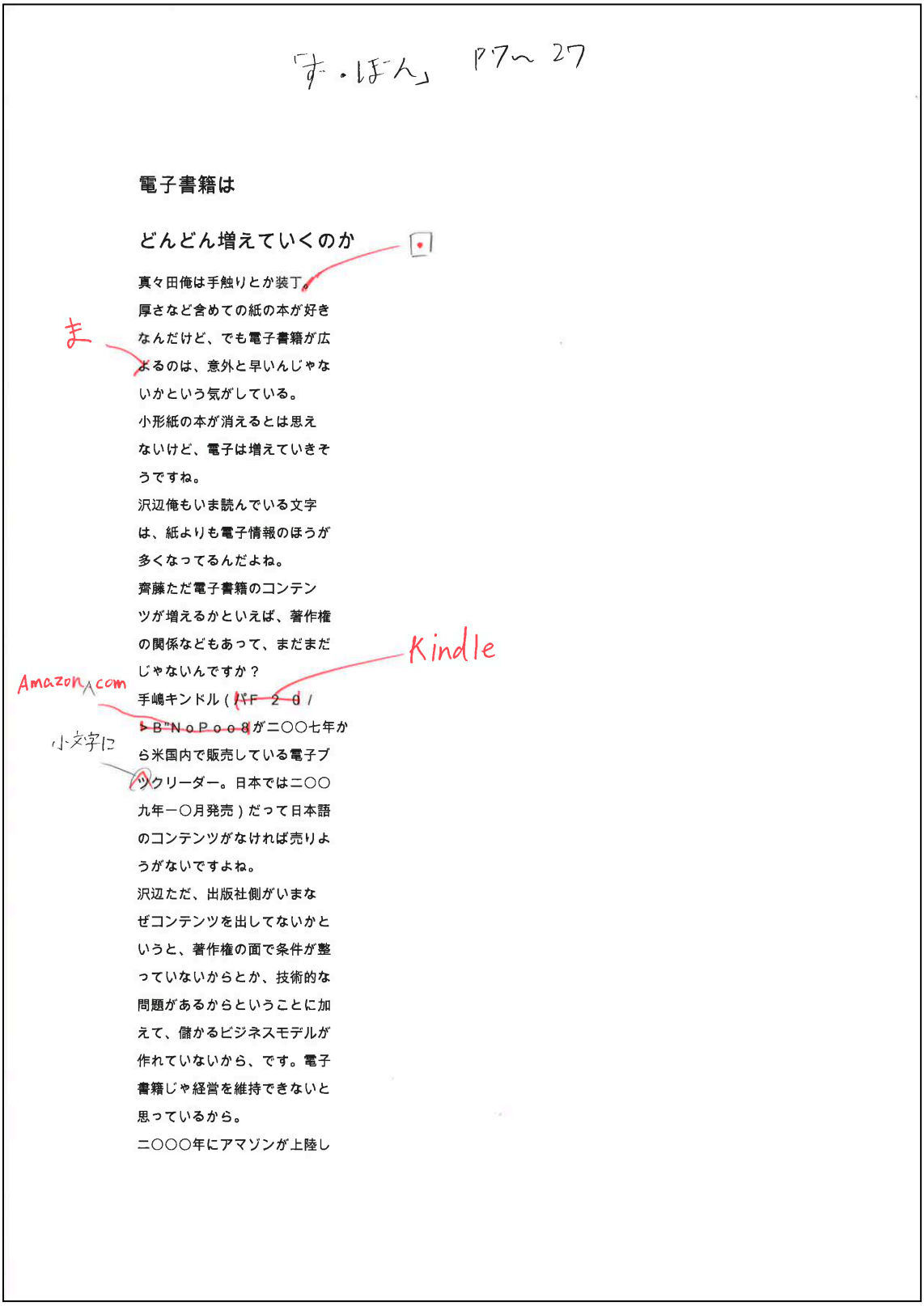
発生した赤字を、「記号」「日本語」「英語」「数字」の別にカウントし、OCR テキストデータの文字数で除して OCR 処理の全体精度を把握する。

照合事例から共通に読み取れるもの、特殊な事例だと思われるものを分析者が読み込み、レポートする。

[別紙 Ⓐ]

【照合結果イメージ】

「ず・ぼん（ポット出版）16 号」P7～



ず・ぼん」 P7～27

電子書籍は

どんどん増えていくのか

真々田俺は手触りとか装丁、
厚さなど含めての紙の本が好き
なんだけど、でも電子書籍が広
まるのは、意外と早いんじゃな
いかという気がしている。
小形紙の本が消えるとは思え
ないけど、電子は増えていきそ
うですね。
沢辺俺もいま読んでいる文字
は、紙よりも電子情報のほうが
多くなってるんだよね。
齊藤ただ電子書籍のコンテン
ツが増えるかといえば、著作権
の関係などもあって、まだまだ
じゃないんですか？
手嶋キンドル（PF-2-0/
ｰB"NoPoo8）が二〇〇七年か
ら米国内で販売している電子ブ
ックリーダー。日本では二〇〇
九年一〇月発売）だって日本語
のコンテンツがなければ売りよ
うがないですよね。
沢辺ただ、出版社側がいまな
ぜコンテンツを出してないかと
いうと、著作権の面で条件が整
っていないからとか、技術的な
問題があるからということに加
えて、儲かるビジネスモデルが
作れていないから、です。電子
書籍じゃ経営を維持できないと
思っているから。
二〇〇〇年にアマゾンが上陸し

Kindle

Amazon.com

小文字に

[別紙 Ⓐ]

『日本の出版流通における書誌情報・物流情報のデジタル化とその歴史的意義（ポット出版）』

「日本の出版流通における～」P. 19～40
（P.24, 25, 30 は表のため除く）

第１章日本図書コードおよびＩＳＢＮ導入問題とは何か

１　日本図書コードおよびＩＳＢＮ導入問題の概要

ここでは日本図書コードおよびＩＳＢＮ導入問題とは何かを示し、その日本の出版流通研究における位置を明らかにし、本研究の研究方法を述べる。

日本で出版される書籍にＩＳＢＮ（国際標準図書番号）が表示されるようになったのは１９８１年１月からである。正確に言えばＩＳＢＮをキーコードとする「日本図書コード」が導入されたのである。日本図書コードとは、

１０桁のＩＳＢＮ（国際標準図書番号）、「販売対象」「発行形態」「内容」を示すコードと呼ばれる４桁の図書分類コード、価格によって成るコード体系である。すなわち日本図書コード＝ＩＳＢＮ＋図書分類コード＋価格コードであり、ＩＳＢＮ＝国別記号＋出版者記号＋書名記号＋チェックデジット（チェック数字）によって構成されている。本書では日本図書コードとＩＳＢＮは明確に区別して論じるが、一般的には混同されて用いられることが多い。

日本図書コードおよびＩＳＢＮの導入に際しては、この構想が発表された当初より出版流通対策協議会（以下、流対協）、図書コードの問題を考える会、日本出版労働組合連合会（以下、出版労連）によって「本の総

19 [illegible]

背番号制」であるという反対運動が起こった。また、日本図書コードおよびＩＳＢＮの導入後は、図書館労働者交流会、図書館事業基本法に反対する会、図書館を考える会、学術情報システムを考える会（１９９０年４月より巨大情報システムを考える会に改称）、全国一般労働組合大阪府本部旭屋書店支部（以下、全国一般労組旭屋書店支部）などの団体がさまざまな形でその導入を批判している。そうした批判は日本図書コードおよびＩＳＢＮそのものだけでなく、出版物のコード化がもたらす流通合理化のさまざまな位相におよぶため、反対運動の主張は多様で複雑である。

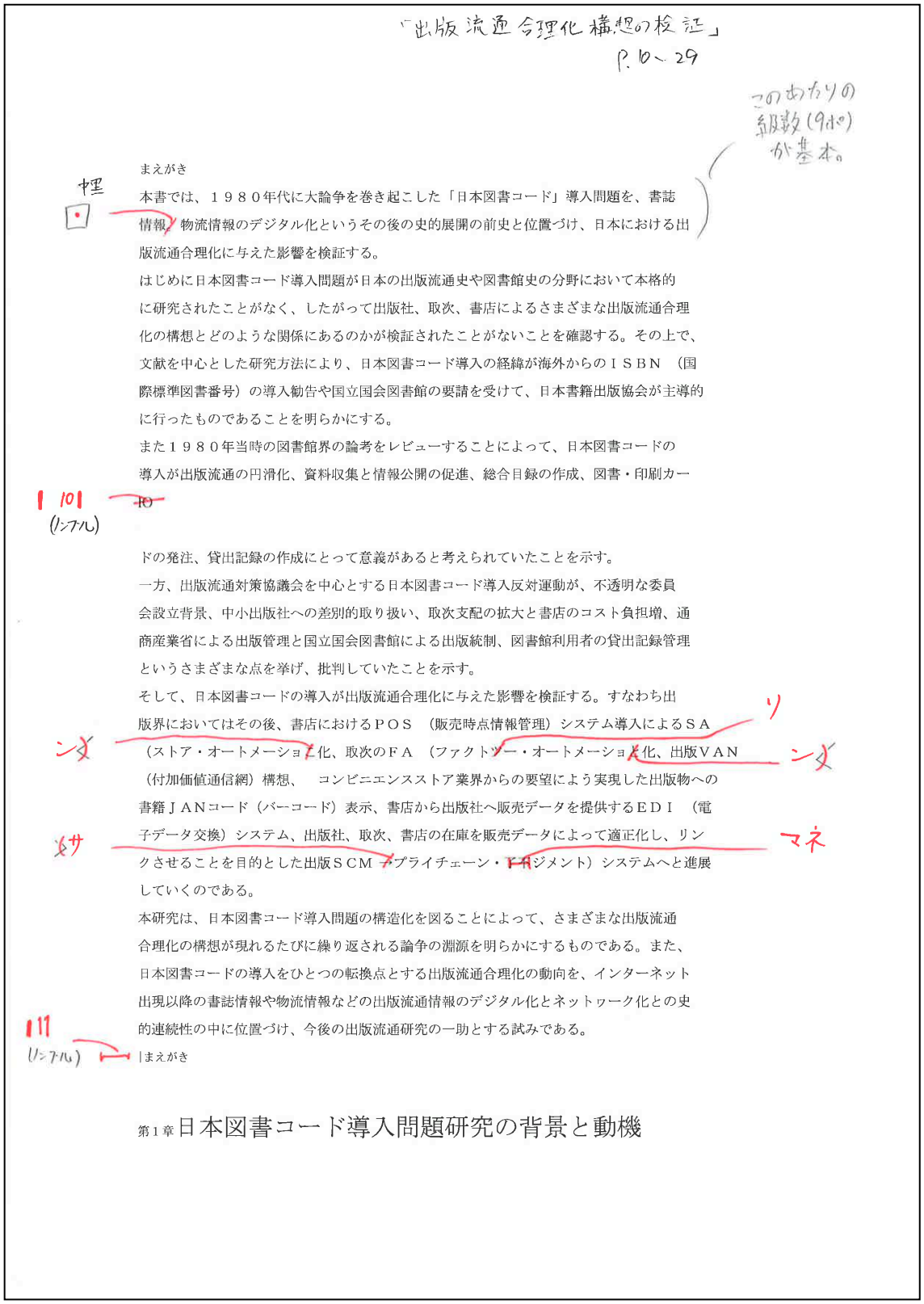
２　取次におけるコンピュータ導入の軌跡

日本の取次に初めてコンピュータが導入されたのは１９６４年である。大手取次である東京出版販売（１９９２年よりトーハンに社名変更）は１９６４年５月に初期のコンピュータ「ＵＮＩＶＡＣ１００４」を導入し、雑誌送品票の作成に利用した。その後、１９６８年７月には「ＨＩＴＡＣ８３００」による新刊書籍の送品票作成業務の機械化を実施、１９６９年１０月にはコンピュータ利用の高度化を図るために業界初の大型コンピュータ「ＨＩＴＡＣ８４００」を導入した。この大型コンピュータの導入について東京出版販売の社史は次のように記述している。

この導入の目的は、①事務の省力化・合理化をさらに推し進め事務コストの軽減を図る②経営管理の機械化を進め、コンピュータの高度利用による情報の有効活用を図る③業界統一書籍コードの実施（１９７０年１月＝引用者注）に伴う書籍業務の拡充、発展を目指す④近代的な出版流通センターの完成を目指す全社的な目標に対応する情報処理体制を確立するなどであった。

[別紙 Ⓐ]

『出版流通合理化構想の検証（ポット出版）』



「出版流通合理化構想の検証」
P.10〜29

まえがき

本書では、１９８０年代に大論争を巻き起こした「日本図書コード」導入問題を、書誌情報、物流情報のデジタル化というその後の史的展開の前史と位置づけ、日本における出版流通合理化に与えた影響を検証する。

はじめに日本図書コード導入問題が日本の出版流通史や図書館史の分野において本格的に研究されたことがなく、したがって出版社、取次、書店によるさまざまな出版流通合理化の構想とどのような関係にあるのかが検証されたことがないことを確認する。その上で、文献を中心とした研究方法により、日本図書コード導入の経緯が海外からのＩＳＢＮ（国際標準図書番号）の導入勧告や国立国会図書館の要請を受けて、日本書籍出版協会が主導的に行ったものであることを明らかにする。

また１９８０年当時の図書館界の論考をレビューすることによって、日本図書コードの導入が出版流通の円滑化、資料収集と情報公開の促進、総合目録の作成、図書・印刷カードの発注、貸出記録の作成にとって意義があると考えられていたことを示す。

一方、出版流通対策協議会を中心とする日本図書コード導入反対運動が、不透明な委員会設立背景、中小出版社への差別的取り扱い、取次支配の拡大と書店のコスト負担増、通商産業省による出版管理と国立国会図書館による出版統制、図書館利用者の貸出記録管理というさまざまな点を挙げ、批判していたことを示す。

そして、日本図書コードの導入が出版流通合理化に与えた影響を検証する。すなわち出版界においてはその後、書店におけるＰＯＳ（販売時点情報管理）システム導入によるＳＡ（ストア・オートメーション）化、取次のＦＡ（ファクトリー・オートメーション）化、出版ＶＡＮ（付加価値通信網）構想、コンビニエンスストア業界からの要望により実現した出版物への書籍ＪＡＮコード（バーコード）表示、書店から出版社へ販売データを提供するＥＤＩ（電子データ交換）システム、出版社、取次、書店の在庫を販売データによって適正化し、リンクさせることを目的とした出版ＳＣＭ（サプライチェーン・マネジメント）システムへと進展していくのである。

本研究は、日本図書コード導入問題の構造化を図ることによって、さまざまな出版流通合理化の構想が現れるたびに繰り返される論争の淵源を明らかにするものである。また、日本図書コードの導入をひとつの転換点とする出版流通合理化の動向を、インターネット出現以降の書誌情報や物流情報などの出版流通情報のデジタル化とネットワーク化との史的連続性の中に位置づけ、今後の出版流通研究の一助とする試みである。

まえがき

第１章 日本図書コード導入問題研究の背景と動機

［別紙 Ⓐ ］

【照合結果カウント結果・照合精度】

書名	校正文字数	赤字発生字数					
		記号	英字	数字	日本語	赤字合計	赤字合計／校正文字数
デジタルコンテンツをめぐる〜	12,675	140	44	51	265	500	3.9%
出版流通合理化〜	9,946	137	88	74	32	331	3.3%
どすこい出版流通〜	14,291	73	117	30	75	295	2.1%
石塚さん、〜	12,999	95	0	65	66	226	1.7%
日本の出版流通における〜	16,936	126	96	114	128	464	2.7%
本の現場	10,335	22	18	68	61	169	1.6%
図書館の近代	11,441	348	4	86	656	1,094	9.6%
図書館という軌跡	17,590	113	31	53	252	449	2.6%
ず・ぼん	23,831	121	366	72	146	705	3.0%
千代田図書館とはなにか	15,025	54	124	37	26	241	1.6%
出版時評	13,143	146	5	14	439	604	4.6%
計	158,212	1,375	893	664	2,146	5,078	3.2%

1 ページあたりおよそ 500〜900 字の単行本（「ず・ぼん」は 4 段組雑誌なので 1 ページあたり約 1200 字）11 冊の本文冒頭 20 ページの OCR テキストデータの文字数合計は 158,212 字。

赤字発生件数は 5,078 字。
約 97%の精度で OCR テキストデータは本文を再現している。

赤字合計 5,078 字の内訳：

日本語（ひらがな、カタカナ、漢字）で発生した赤字は 2,146 字（赤字中約 42.3%）。
記号（パーレン・ナカグロ・句読点等）で発生した赤字は 1,375 字（赤字中約 27.1%）。
英字で発生した赤字は 893 字（赤字中約 17.6%）。
数字で発生した赤字は 664 字（赤字中約 13.1%）。
日本語の文字数の割合が圧倒的に高いことを勘案すると、記号・英字・数字での精度が相対的に悪いことが示されている。
（※必要に応じて各文字比率を厳密にカウントすること要検討）

［別紙 Ⓐ ］

【OCR テキストデータにおける一般的傾向】

記号では、パーレン（ ）《 》【 】 ・ ナカグロを誤読する事例が多い。
パーレンで囲んだ文章も誤読する事例が少なくない。
その影響で前後の通常文章にも誤読が発生するケースがある。
二重カギ『』も少なからず誤読される。
英字では、縦組中に横組で表記された英字は全く読み取らない。
数字は、1 を 7 と読み込むなどの事例が散見されるが、ノンブルの読み込みに失敗する場合も多い。（※目次ノンブルと本文ノンブルのリレーションを生成させるような場合の阻害要因となりうる）
表組に誤読が多い。
装丁デザインと一緒に組んだ文字は読めない。
級数の変化があると誤読が増加する。（※頻出級数をターゲットにして OCR 処理のチューニングを行っているためか？）

【書籍によってばらつきはあるが頻出する傾向】

特定の日本語文字に誤読が見られる。
（ぎ、ざ、しょう、り、て、ほ、は、に、せ、じ、聞、問、間、目、日、口、？）
ナンバリングの数字に誤読が多い。
(1)、 (2)、(3)
二桁数字の表示に誤読が多い
（年代表記の「６０年」「７０年」など。「２３」「４５」「９９」「１４」など）

【今後の課題】

本レポートにおいては、初回、時間的などの諸制約により OCR テキストデータの本文照合作業とそれによる OCR テキストデータの精度の明確化がメインとなったが、必要に応じて以下の課題等について調査を検討しうる。
本レポートの使用目的に鑑み、校正者照合済赤字を参考に、以降調査の方向性を定めてより有益性の高いレポートとしたい。

《調査課題例》

・日本語、記号、英字、数字各種文字について個別の精度検証
・発生する誤読傾向（ぎ→ざ、三→二、＜→く etc.）の調査
・表組の多いページ（巻末数字データ等）、目次などの特殊ページ、見出し部分などの精度検証
・別種工程により生成されたテキストデータの精度検証　etc.

OCR 照合分析：株式会社ゼロメガ