美しいPDFファイルを作る ： 機関リポジトリのための冊子体電子化手順

星子, 奈美
九州大学附属図書館

出版情報：九州地区大学図書館協議会誌. 51, pp.10-13, 2009-02. 九州地区大学図書館協議会
バージョン：
権利関係：

寄　稿　論　文

美しいPDFファイルを作る

－機関リポジトリのための冊子体電子化手順－

星　子　奈　美

１．はじめに

機関リポジトリに登録される研究成果物は、すでに電子ファイルが存在するものばかりではなく、冊子体から電子ファイルの作成を行う必要がある場合も多い。「九州大学学術情報リポジトリ」では、研究紀要をはじめとする学内刊行物の包括的な登録を推進しているが、刊行年の古い冊子については、冊子をスキャンして電子ファイルに変換するという作業過程を経る。作業量は決して少なくないが、年月を経て傷みの激しい冊子体を、電子ファイルとして新たに生まれ変わらせ、オンライン上で広く利用に供することができるというのは、機関リポジトリを運営する大きな意義の一つである。

九州大学では、冊子体の電子化を効率的におこなうため、試行錯誤を重ね、作業手順を整理した。以下にその手順を詳述する。

２．使用機器およびソフトウェア

電子化にあたっては、下記の機器およびソフトウェアを使用している。

スキャナ

DocuScan C4250…裁断した冊子からPDFファイルを作成する際に使用する、コピーとスキャナの複合機である。後述の「ApeosWare Flow Service」と連携している。

plustek OpticBook 3600…冊子を裁断せずにスキャンする場合に使用する。

OCR処理用ソフトウェア

ApeosWare Flow Service…PDFファイルに透明テキストを付与する際に使用する。

画像加工用ソフトウェア

Adobe Photoshop…スキャンした画像を補正する際に使用する。

Adobe Acrobat…複数のPDFファイルの結合や日本語・英語以外のOCR処理の際に利用する。

３．電子化の手順Ⅰ：冊子を裁断する場合

冊子の残部が複数現存しているなどの理由で冊子を解体しても差し支えがない場合は、裁断機で背の部分を切り落とし、１枚ごとに分解した後スキャンする。

①　裁断作業

裁断機で冊子の背を裁ち落とす。

②　スキャン作業

裁断した冊子をDocuScan C4250でスキャンする。コピーを取る際と同じ要領で、自動紙送り装置へスキャン原稿を挿入することにより両面スキャンが可能である。機関リポジトリ登録の利便性を考慮し、スキャンは１論文単位でおこなっている。

スキャンに際しては、読み取りサイズを自動設定にせず、実寸のページサイズより小さめに指定する。これは、PDFファイルの端に黒線が入ってしまう現象を防ぐためである（図１）。

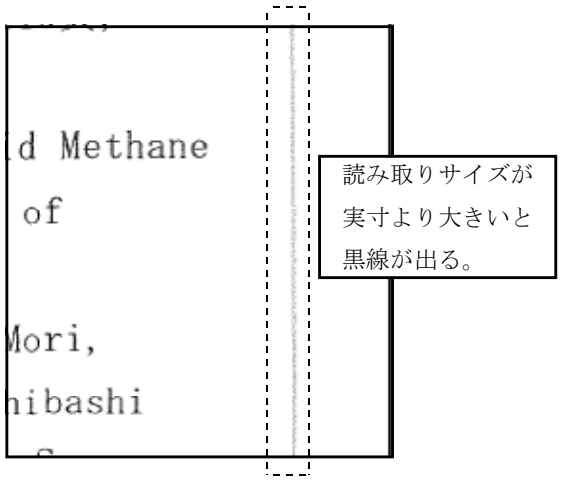

図１　PDFファイルに入った黒線

また、スキャン時のカラーモードはモノクロ２階調、解像度は400dpiを基本とする。原稿にカラーページやモノクロ写真などがある場合は、

フルカラーもしくはグレースケールで該当ページのみをスキャンし、後にAdobe Acrobatでファイルを結合する。

スキャンした画像は、PDFファイルとして、ネットワーク接続したデータストレージに転送・保存される。本学では、「日本語」「英語」「日本語・英語混在」「その他（OCR変換なし)」の４つのデータフォルダを作成し、スキャンする論文の使用言語に応じて、保存先フォルダを指定している。

③　OCR処理

ApeosWare Flow Serviceをインストールしたコンピュータ上で、同ソフトウェアを起動すると、データストレージの「日本語」「英語」「日本語・英語混在」のフォルダに保存されたPDFファイルに透明テキストを付与する処理が開始される。

言語別にフォルダを分けるのは、OCRの精度を高めるため、適用する文字認識ルールをフォルダごとに設定しているからである。日本語・英語以外の場合は、ApeosWare Flow Serviceではなく、Adobe Acrobatのテキスト認識機能を用いてOCR処理をおこなうため、スキャンデータの保存時に「その他（OCR変換なし)」のフォルダを選択する。

以上の手順を経て作成されたPDFファイルが、機関リポジトリへ登録される。

４．電子化の手順Ⅱ：冊子を裁断しない場合

冊子の残部がない場合や、冊子を論文著者に返却する必要がある場合は、冊子を裁断せず、plustek OpticBook 3600を用いて１ページずつスキャンした後、Adobe Photoshopを用いてPDFファイルを作成している。

①　スキャン作業

plustek OpticBook 3600では、図のように冊子を１ページずつスキャンする（図２)。スキャナの角まで画像の読み取り面があるため、ページの中央部分までスキャン可能で、見開きでスキャンした場合に生じるような影ができにくい。また、冊子を大きく開く必要がないため、冊子に対する負担も少ない。

図２　plustek OpticBook 3600でのスキャン

②　ファイル加工

ⅰ．Adobe Photoshop起動

↓

ⅱ．カラーモード変更　カラーモードを「モノクロ２階調」から「グレースケール」に変更する。これは、次工程での画像の回転に際し、任意の角度を指定できるようにするためである。

↓

ⅲ．画像の回転　スキャン時に生じた画像の傾きを調整する。画像の回転を何度も繰り返すことは、画質の悪化につながるため、回転角度は一度で確定するのが望ましい。この際に便利なのが「ものさしツール」である。画像が水平になるよう、基準となるライン(枠線など)に合わせて「ものさしツール」で線を引き、メニューから「イメージ」→「カンバスの回転」→「角度入力」を選択すると、水平にするために必要な回転角度が自動的に測定される。

↓

ⅳ．切り抜き　原稿の文字部分が画像中央に配置されるよう、周囲の余白を一旦切り抜いて削除する（図３－１)。

↓

ⅴ．画像サイズ設定　画像のサイズを再設定する。メニューから「イメージ」→「カンバスサイズ」を選択し、「基準位置」で現在の画像がカンバスの中心に配置されることを確認する。「幅」および「高さ」に仕上がりサイズを入力し、画

像が冊子の現物と同サイズになるように調整する（図３－２）。

図３－１　周囲の余白を切り抜いた状態

図３－２　画像サイズを再設定した状態

↓

vi．ゴミ消し　原稿が古い時や、シワが多い時、裏面の印字が写っている時などはゴミが増え、OCR処理の精度に影響する。範囲選択して削除、もしくは消しゴムツールで消していく（図４－１、４－２）。

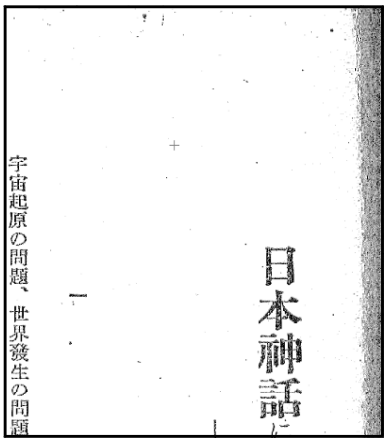

図４－１　ゴミ消し前

図４－２　ゴミ消し後

↓

vii．カラーモード変更　「グレースケール」に変更していたカラーモードを「モノクロ２階調」に戻す。

↓

viii．PDFファイル保存　「ファイル」→「別名で保存」でPDF形式に変換して保存する。

以上の工程で、１ページ分の画像処理が完了するので、各ページ同様に加工していく。同じ工程を繰り返す必要がある場合は、「アクション」機能を利用し、作業手順を記録すると効率がよい。例えば、原稿に汚れが少なく、ゴミ消しの手順が不要な場合には、iv→ｖ→vii→viiiの工程を一連のアクションとして記録しておけば、次回以降はそのアクションを再生するだけで一連の作業が完了する。

③　PDFファイルの結合

１ページずつ加工して作成したPDFファイルを論文単位でひとつのファイルにまとめる。Adobe Acrobatのファイル結合機能を利用する。

④　OCR処理

以後の手順は、冊子を裁断する場合と同じである。

５．おわりに

機関リポジトリにおいて美しいPDFファイルの提供を心がけることは、ファイルをダウンロードして利用する方々の利便性を向上すると同

時に、機関リポジトリの信頼性を高めることになる。そして、信頼度の高い機関リポジトリの運営は、更なる登録促進にもつながると考えられる。今後も作業手順を逐次見直しながら、質の良いPDFファイルの提供に努めたい。

６．謝　　辞

作業手順の整理にあたっては、画像加工の担当スタッフとして従事されていた松嶋寛子氏に多大なご尽力をいただいた。この場を借りて深く感謝申し上げる。

［参考］

九州大学学術情報リポジトリ
https://qir.kyushu-u.ac.jp/

DocuScan C4250/C3200 A
http://www.fujixerox.co.jp/product/docuscan_c4250/

ApeosWare Flow Service
http://www.fujixerox.co.jp/product/aw_flow_service/

plustek OpticBook 3600
http://www.plustek.com/product/book3600.asp

Adobe Photoshop
http://www.adobe.com/jp/products/photoshop/

Adobe Acrobat
http://www.adobe.com/jp/products/acrobat/

ほしこ　なみ
（九州大学附属図書館）