講座◆ 誰でもできる情報管理

第11回

紙媒体からデジタル情報への変換

堀越敬祐^{*1}

[著者抄録] まず最初に、紙に書かれた情報をデジタル情報に変換（電子化）する意義を考えてみる。次に、紙媒体が持つ様々な属性ごとに電子化の注意点を列挙する。最後に、多種多様な電子化データ形式のうち、代表的な形式の特徴とその用途も簡単に論述する。

[著者付与キーワード] 電子化, 紙媒体, ファイル形式, 画像データ, スキャニング

Series◆ Information literacy in your hands

11

Convert from paper format to digital data

HORIKOSHI Keiyu

[Author Abstract] First of all, the meaning of conversion (digitalize) from written on the paper to digital information is described. Next, notes of digitalizing of each various attributes of the paper medium are enumerated. Finally, the feature and the usage of a typical form are easily stated among various digitalization data forms.

[Keywords by Author] digitalization, digitization, paper format, digital file format, graphic data, scanning

^{*1} 国立国会図書館 総務部企画課 電子図書館推進室 (〒100-8924 東京都千代田区永田町1-10-1 Tel. 03(3581)2331 E-mail: horikosi@ndl.go.jp
National Diet Library, Electronic Library Development Office, Planning Division,
Administrative Department
(10-1, Nagata-cho 1-chome, Chiyoda-ku, Tokyo, 100-8924)

1. 電子化の意義

紙媒体を電子化することによって得られるメリットは以下の通りである。

(1) 保管場所の縮小

法律により紙媒体の原本の保存が定められた文書、法的な手続きにおいて紙媒体の原本でないと困る文書、図書館や文書館が所蔵する資料などには効果が薄い。一般的な文書は電子化した後に廃棄することで、保管スペースを大幅に減少させることができる。

(2) 保存性の強化

電子化した情報自体は複製によって劣化しないため、複製物を作製し分散して保管することで、災害から情報を守ることができる。

(3) アクセスの簡便化

電子化した情報をネットワーク上に置くことで、遠隔地からも利用できるようになる。また、モバイルPCに情報を作製しておき、ネットワークの使えない場所で利用するという方法も考えられる。

(4) 共有化の促進

紙媒体だと、誰かがファイルを使っていると、基本的に他の者が使えないが、電子化して共有の設定をしておけば、複数の者が同時に利用できるようになる。

(5) 情報の一元化

最近ではほとんどの事業所で文書の作成にPCを用いていると思う。また、取引先とも電子メールを介したデータのやり取りが一般化している。こうした情報は最初から電子的であるから、その状態のまま蓄積、活用したい。

2. 電子化の実現に向けて

まず決めなければいけないことは、電子化を自前で行うか、外注するかということだろう。判断の材料は、電子化の対象となる文書の種類、現時点での件数と枚数、その増加予測値、そして電子化した文書の利用体制である。利用体制には、誰がどのようにして電子化文書を閲覧するかという

ことだけでなく、新規の電子化、機器やソフトウェアの保守、日常的なバックアップなどの付帯作業も含まれる。また、外注する範囲にも電子化だけの場合や丸投げの場合があるが、利用のことを考えると、対象となる文書の選定と、電子化した文書へのキーワードの付与は自前で行った方がよい。いずれにしても電子化は手間やコストがかかるものであるから、判断材料を集め、業者から見積もりをもらい、必要に応じてトライアルしてみ、手間やコストに見合った効果が得られないと結論づけられたら、見送るのも一つの選択肢である。文書の保管場所に困っているという事情だけならば、貸倉庫の利用やマイクロ化でも代用は可能である。

3. 電子化の対象となる紙媒体

一口に紙媒体といっても様々な種類があり、電子化の方法や電子化情報の形式を決定する大きな要因となる。

3.1. 形態

A4判くらいまでの一枚物はシートフィーダ（給紙装置）を使い、連続してスキャンすることができる。冊子は綴じの部分で断裁して一枚物としてスキャンするのが効率的であるが、解体することになり、綴じに近い部分に書かれた文字が判読しづらくなることがある。巻物などの特殊な形態の場合はスキャナを用いず、銀塩カメラで撮影して、フィルムをスキャンするのが普通である。

3.2. 紙の状態

劣化した紙、裏写りが顕著な薄い紙、青焼きコピーなどは通常の電子化では利用に耐えうるデータの作成が困難であるか、下手をすると電子化が不可能な場合がある。電子化の方法を工夫したり、データをソフトウェアで修正する必要がある。

3.3. 大きさ

電子化の対象となる紙媒体は一般的にA4判やB5判であり、大抵のスキャナもこれら大きさに対応している。これ以外の大きさの紙媒体で電子化される機会が多いのが、小さいものと名刺、大きいものと地図や設計図面であろう。名刺は専用のスキャナがあり、地図や設計図はドラムスキャナを用いる方法と、大判カメラで撮影してフィルムをスキャナにかける方法がある。

3.4. 文書の内容

用途にもよるが、内容が分かればよいものと、可能な限り原本に忠実な再現が必要なものがある。前者は事務文書全般が当てはまる。後者は古典籍などの貴重資料の、研究者向けの電子化が挙げられる。

3.5. 色の有無

カラーの文書は、色の再現が重要になることがある。特にカラー写真など自然色の文書を電子化する際には、カラーマッチングの設定をきっちり行い、悲惨なデータができないように注意しなければならない。やむなくモノクロでスキャンするのであれば、2階調ではなく、グレースケールにした方がデータの品質面で無難であろう。

3.6. 図版の有無

文字だけで構成される文書の電子化は、白黒のみの2階調でスキャンするのがポピュラーであるが、そこに写真や線画でない絵が含まれていると、その部分は何が何だか分からない画像が生成される。対策としては、やはりグレースケールでスキャンするしかない。

3.7. 文字の状態

手書き（筆、ペン、鉛筆など）は文字に濃淡があり、個別に調整しないと画像中の文字が読めないことがある。また、古書籍の活字や古いワープロで作成されたドットの粗い字体は、OCRによる文字認識が困難である。

3.8. マイクロフィルム

マイクロフィルムは、文書の保存媒体として長い実績がある。大量の文書を扱う機関では、すでに紙媒体から変換し、蓄積しているところも多い。ただ、マイクロフィルムは同時に複数の利用者がアクセスできない点が不便なため、マイクロフィルムを電子化することも積極的に行われている¹⁾²⁾。また、大きさや状態により直接スキャンできない文書を、マイクロ撮影したうえでフィルムスキャナにかける電子化もある。問題は、既存のマイクロフィルムが2値で撮影されていて、マイクロリーダーで見るとそれほど違和感がないものの、電子化すると元がカラーや図版の文書は、前述の通り役に立たないデータとなることである。画像編集ソフトウェアを用いて擬似的なグレースケールに変換することも可能であるが、元からグレースケールの画像とは比べるべくもない。

4. 電子化データの形式

電子化によってできあがるデータには、テキストデータとイメージ（画像）データの2種類がある。イメージデータの形式は、ビットマップとベクトルに大別されるが、紙媒体の電子化では、ベクトル画像データを生成する事例はあまり聞いたことがない。

4.1. テキスト化

紙媒体をスキャナで画像データに変換し、その画像データ上にある文字をOCRソフトによって1文字ずつ認識してテキスト出力するものである³⁾。テキストだとファイルのサイズが小さい、キーワードを付与せずとも全文検索で利用ができるといった利点があるが、現在のOCRソフトでは100%の識字率が困難であり、必ず人が確認と修正を行わなければならない、イメージ化に比べて手間とコストがかかりすぎる欠点がある。したがって、採用するとすれば、絵や写真といった図版を含まない、印刷状態が良好で定型的な書式の文書や、テキストでなければ利用する意味がない名刺

などにとどめておくべきであろう。名刺のテキスト化については、パーソナル用途のスキナとソフトウェアを組み合わせた製品が市販されている（使い勝手は不明）。また、テキスト化をした後のデータ保存形式は、レイアウトをどの程度まで維持するか、どのような利用をするかで様々な選択肢が存在する。主な形式としては、プレーンテキスト、HTML、XML、MS-Wordなどのワープロ形式、PDF形式などが考えられる。

4.2. イメージ

書かれている文字を文字として認識せず、とにかく紙媒体を「見たまま」のデータとして扱う形式である。テキスト化よりも手間とコストが安くすむため、大量の文書を電子化する場合は、ほとんどこの形式が採用されている。イメージデータといってもその中で形式が無数にあり、紙媒体の電子化では主に以下の形式が用いられている。

4.2.1. TIFF形式⁴⁾⁵⁾

ほとんどの画像処理ソフトで扱うことができる。スキナで読み込んだイメージは、特に指定しないと大抵はTIFF形式で保存される。その一方で、実際にはリビジョンや色情報の扱い、圧縮方式で様々なバリエーションがあり、混乱を来すことがある。また、ブラウザで直接表示することができないので、インターネット上で利用されることはあまりなく、組織内部で運用される文書管理システム内や、利用を前提としない完全な保存用途で使われることが多い。

モノクロの文書であれば、FAXの送受信で使われるMMR圧縮のTIFF形式が非常に有効である。データで白か黒かの2値で記録されるため、ファイルサイズが他形式と比べても圧倒的に小さい。複数のイメージを一つのファイルで管理するマルチページTIFFというバリエーションもある。NDLでは、衆参と共同で構築した国会会議録の検索システムで、検索結果の印刷用イメージ表示用（一部は検索結果そのもの）として用いられている。また、国際子ども図書館においても、館内マルチメディアシステム用のデータとして、児童書の本

文ページをMMR圧縮のTIFF形式で電子化している。

カラーの文書や多階調でないと表示に問題がある文書は、カラーモードやグレースケールのTIFF形式を用いる。ファイルサイズが大きいので、一般的にはLZW圧縮を使用する。LZW圧縮は、圧縮によってイメージの劣化が生じない、いわゆる可逆圧縮である。この圧縮で注意したいのは、他形式から変換したり、表示したりするソフトウェアを開発する際は、圧縮アルゴリズムのライセンス供与を米UNISYS社から受ける必要があることである⁶⁾。もちろん、市販のソフトウェアでは、ライセンス供与のうえで作られているため、これらを使用する分には何ら問題は発生しない。NDLでは、ホームページの「デジタル貴重書展⁷⁾」で公開している「伊能図」の元画像がこのTIFF形式である。8×10サイズの大判フィルムを1,200dpiの読みとり解像度でスキャンしたデータで、圧縮してあるとはいえ、1ファイルで200メガバイトくらいある。

4.2.2. JPEG形式⁸⁾

インターネット、とりわけWWW上で扱われる画像形式としては標準的であり、最も普及している。圧縮は、劣化を伴う不可逆圧縮方式であるが、画質と引き替えに高圧縮によりカラーまたはグレースケールのTIFFとは比べものにならないほど小さいファイルを作成することができる。ただし、あまり圧縮をしすぎると汚い画像になり、文字は滲んだような感じで非常に読みづらくなる。写真を含むカラーの文書を、インターネットで公開する目的ならば有効な選択肢であろう。なお、JPEG形式の保存には、プログレッシブJPEGなどいくつかオプションする。通常はオプション指定は不要である（ファイルサイズが少し大きくなる）。

4.2.3. GIF形式

この形式もほとんどのWWWブラウザでサポートされているため、インターネットで多く使われている。JPEG形式との違いは、インデックスカラーと呼ばれる256で色を扱うため、JPEG形式

のように自然色の表示には向いていないが、その中に収まる色数のデータであれば、JPEG形式よりもきれいな画像で表示できる。LZW圧縮を使用している。

4.2.4. PDF形式⁹⁾

正確には、画像データの形式ではなく、ワープロや表計算などのテキストデータからも作成可能な、本当の意味での文書形式である。Adobe社が独自に開発したファイル形式であるが、表示用のソフトウェア「Acrobat Reader」を無料配布し、インターネット上で配布する文書のみならず、CD-ROMなどの媒体に固定して配布する文書においても、非常に多く採用され普及している。とにかく、プリンタに印刷できるデータであれば、ページ数や用紙サイズに関わりなく、他のあらゆる形式から変換が可能である。PDF形式をイメージのカテゴリに入れたのは、インターネットでの配布に不適当なTIFF形式データを、PDF形式に変換することで表示に関わる問題を取り除くことができるのが理由である。TIFFを変換したPDFは、表示ソフトの機能である文字列検索や、表示している文字列の範囲指定ができないといった制限があるが、ディスプレイ上の表示、プリント出力では元データの画質に準じた、それなりの品質が得られる。

4.3. その他の不適な形式

Windowsにおけるbmp、Macにおけるpictといった画像形式は、そのOSでは標準的に扱えるが、他OSでは別途ソフトウェアをセットアップしないと表示できないし、ファイルのサイズとしても圧縮していないため大きくなりがちであるから、文書を電子化する形式としては不向きである。

4.4. その他の有効と思われる形式

4.4.1. PhotoCD¹⁰⁾

Kodak社独自の画像形式。そもそも媒体としてのCD-ROMの形式からして独自である。複数のピクセル解像度のデータを内包し、開くときに指

定することができる。その名の通り、写真の電子化には向いているが、扱える画像の最大ピクセル解像度が4,096×6,144（プロPhotoCD）までという制限がある。

4.4.2. FlashPix¹¹⁾

複数の企業が共同で開発した画像形式。PhotoCDと同様に、複数のピクセル解像度を内包し、プラグインを組み込むことで、ブラウザ上での表示が可能となる。ただし、利用者が指定する解像度で表示できるようにWeb上から配信するには、専用のサーバが必要となる。

4.4.3. PNG¹²⁾

可逆圧縮により画像の劣化がなく、ファイルサイズもGIFと同程度になる。比較的新しいバージョンのブラウザでは標準で表示できる。W3CがWeb上での使用を薦めているが、あまり普及している気配はない。

5. 電子化の方法

通常はスキャナで紙媒体を光学的に読みとり、電子化を行う。通常は、というのは、3章の電子化の対象となる紙媒体で示したように、フィルムを経由して電子化する方法もあるからである。また、最近ではデジタルカメラの性能向上が著しく、パソコンやプリンタと組み合わせて使うケースが増えている。また、一眼レフカメラの代替となる高性能な機種も登場し、報道現場などで活用されているという。そこで考えられるのは、マイクロカメラの代用としてデジタルカメラを使えないかということで、実現できるなら、マイクロ撮影の要領で冊子形態の文書や破損のおそれのある文書を電子化するコストが削減可能となる。NDLでは、BBCCとの協力事業「世界の中のニッポン」の一部コンテンツ作成で試験的にデジタルカメラを使用している。どの画像がデジタルカメラで撮影したものか、興味のある方は「世界の中のニッポン」ホームページ¹³⁾にアクセスし、探してみていただきたい。

～～～ 参 考 文 献 ～～～

- 1) 金澤勇二. 図書館員のためのステップアップ
専門講座(32) 媒体変換—マイクロ化 デジ
タル化(1). 図書館雑誌. Vol.94, No.8, 2000,
p.562-564
- 2) 金澤勇二. 図書館員のためのステップアップ
専門講座(33) 媒体変換—マイクロ化 デジ
タル化(2). 図書館雑誌. Vol.94, No.9, 2000,
p.723-725
- 3) 山岡正輝. 情報の電子化技術—入門から応
用まで(7) 印刷物の電子化技法—OCRとデジ
タイザの活用. 情報管理. Vol.42, No.8, 1999,
p.694-701
- 4) TIFF Software. (<http://www.libtiff.org/>)
- 5) Niles Ritter. The Unofficial TIFF Home
Page. ([http://home.earthlink.net/~ritter/
tiff/](http://home.earthlink.net/~ritter/tiff/))
- 6) Unisys Corporate. LZW Patent Informa-
tion—Japanese Version. (<http://www.uni>
[sys.com/unisys/lzw/lzwfaq_j.html](http://www.unisys.com/unisys/lzw/lzwfaq_j.html))
- 7) 国立国会図書館. デジタル貴重書展. ([http:
//www.ndl.go.jp/exhibit/50/index.html](http://www.ndl.go.jp/exhibit/50/index.html))
- 8) Joint Bi-level Image experts Group. JPEG
JBIG. (<http://www.jpeg.org/>)
- 9) アドビ(株) Adobe PDF. ([http://www.adobe.
co.jp/products/acrobat/adobepdf.html](http://www.adobe.co.jp/products/acrobat/adobepdf.html))
- 10) KODAK DIAMIC フォトCD. ([http://www.
kodak-imagex.co.jp/d_s/photocd/photocd.
html](http://www.kodak-imagex.co.jp/d_s/photocd/photocd.html))
- 11) Digital Imaging Group. Digital Imaging
Group. (<http://www.dig.gr.jp/>)
- 12) The World Wide Web Consortium. Porta-
ble Network Graphics. ([http://www.w3.org/
Graphics/PNG/](http://www.w3.org/Graphics/PNG/))
- 13) 国立国会図書館 世界の中のニッポン.
(<http://nippon.ndl.go.jp/>)