紙文書の画像化に関する報告

A report on digitalized documents by a scanning device

栗 原 和 夫
Kazuo Kurihara

要　旨

2010 年度経営学部情報管理学科の卒論「紙文書の画像化と検索に関する一考察」[文献 1]で研究された紙文書の画像化手順に従って、朝日大学の健康管理センターに保管されているすべての学生健康管理票を画像化した結果を報告する。対象とした学生健康管理票は、歯学部が 9 分冊、経営学部が 5 分冊、法学部が 3 分冊、大学院と歯科衛生士専門学校が各々1 分冊で、全部で 19 分冊であった。スキャナーは A3 版までスキャン可能な、フラットベッドスキャナ、EPSONSCAN の ES-6000HS を使用した。対象とした学生健康診断票は最大のものが縦 257mm 横 364mm の厚手の台紙で、毎年の定期健康診断結果を1枚に 6 年分が貼りこめるようになっている。1 分冊におよそ 200 枚の台紙が綴じられており、文献 1 では、CD1 枚に 38 冊収納できると見積もっている。対象とした 19 分冊は、目次や索引も付けて確かにCD1 枚に収める事ができた。2011 年 3 月 11 日の地震と津波の被害により、自治体に紙で保管されている公文書が津波で流され復元が困難になっていると聞く。今、紙文書の画像化がよりいっそう求められるようになって来ているのではないだろうか。さらにデジタル文書やデジタル化文書であっても、破壊に備えるべく、地域的に離れたサーバに2重保管する等信頼性を高める動きが出て来ている。本報告は、簡単な少量の例に過ぎないが、コストも含めて、紙文書の画像化を検討する判断材料として提示した。

1．はじめに

紙を主体として管理している文書は、地震や津波で喪失すると、復元が非常に困難である。またある時期までは紙文書で管理されていた国民年金関連書類は、自治体ごとに管理が統一されていなかったこともあり、現在もなお納めた保険料の確認照合がなされている。

紙で管理されている文書を、少なくとも画像化し、デジタル化文書にして行くのは、管理上からも重要な作業である。法律で紙での保存が義務付けられている文書はやむをえないとしても、組織内にあるそれ以外の紙文書は、画像化して元の文書を廃棄し、デジタル化文書として保管し、検索キーを付けて管理しやすくすることはもっと考慮されても良い。

1998 年電子帳簿保存法が施行され、もともと紙文書の存在しない取引、つまりネットワーク上のオンライン取引などを考慮し、帳簿書類の保存方法等に関して特例が設けられた。また 2005 年

には e-文書法が施行され、財務税務関係書類に関して、スキャナーで読み込んだ画像ファイルを一定の条件を満たせば原本とみなすことができ、紙の原本を廃棄することができるようになった。

文献 2 によれば、電子データの作成・保存における課題として「見読性」「完全性」「機密性」「検索性」などの確保が必要であるが、対象とする文書によってそれらの内容や技術要件は異なり、文書に応じた個別対応が必要であると述べている。

見読性は、作成した文書を表示・印刷し内容を確認できることで、スキャナーの性能に関わってくる。完全性は、文書の作成者が特定できることや紙文書がその日付以前に存在し現在まで文書の改ざんがなく保管されていることを保証するもので、これには、第 3 者の認証局による電子署名や認定業者によるタイムスタンプが必要である。機密性は文書の操作や保管に関することであり、操作者に対するアクセス制御やファイル管理技術を必要とする。検索性は、必要に応じて求める文書をいくつかのキーワードで探し出せることである。少なくとも文書を特定する 1 対1のキーをすべての文書に付け、このキーで検索できるようにしなければならない。

2. デジタル化対象の学生健康診断票

今回デジタル化の対象とする学生健康診断票の各ページは、1人の学生に対して図1のような台紙である。1人の学生のある年のデータは図1の黒枠に示す部分で、1 年ずつ切り離されて台紙に貼りつけられている。この台紙には 6 年分のデータを貼ることができる。

図1 学生健康診断票の例

台紙の右上の数字がこの学生の学籍番号で、台紙と1対 1 に対応するコードとして利用できる。また学生の健康状況に応じて精密検査が行われ、図2のような診断書が台紙の裏に貼り込まれることがある。

診　断　書

氏　名　[redacted]　殿

生年月日　昭和 38年 3月 28日生

住　所　本巣郡穂積町穂積中屋

1) 現学的所見

i) 身長 167.6 cm　体重 74.0 kg　胸囲 93.0 cm

ii) 血圧 134/86

iii) 所見　特記すべき所見なし

2) 胸部 X 線

所見　異常なし

3) 尿検査

	Hema Combistix 検査
潜血	− + ++ +++
蛋白	− ± + ++ +++ ++++
糖	− + ++ +++
pH	5 6 7 8 9
ウロビリノーゲン	− + ++ +++

上記のごとく診断します

昭和 59年 6月 23日

岐阜県本巣郡穂積町大字穂積1851番地の1
岐阜歯科大学附属病院内科

医師　宮本洋

図2　再検査の診断票の例

修学年限が 6 年の歯学部であれば、通常 6 年で卒業するので、一人分1枚の台紙で間に合うが、精密検査や留年があれば、一人分が 2 枚以上になることもある。このことから、台紙と1対1対応するコードに関しては、学籍番号にさらに記号を付加するなどの対策が必要である。

3.　見読性

e−文書法においては、可視性を確保するための要件として「重要項目の検索機能とディスプレイ、プリンタの備え付け」が財務省令で要請されており、そのための電子計算機処理システムの要件として、「解像度が、日本工業規格 z60164.1.1 に規定する一般文書の変換時の解像度である 1mm あたり 8 ドット(200dpi)以上で読み取るものであること。赤色、緑色および青色の階調がそれぞれ 256 階調（1677 万色）以上で読み取るものであること。」が規定されている。

解像度が上がれば、読み取った画像ファイルのサイズは大きくなる。国税関係書類として保管するのでなければ、見読性を保ちながら、画像ファイルサイズをなるべく小さくすることにより、保管サイズの縮小や検索時間の短縮を期待できる。この点を考慮して、デジタル化対象をいくつかの解

像度で読み取って、見読性を保った解像度を設定することができる。

文献1では、この学生健康診断票に対して、いくつかの解像度を試した結果、見読性を保つとみなされる解像度の中の最小のものとして、解像度 96dpi を、また圧縮率として 50%を提案している。この解像度でスキャンした PDF ファイルを拡大してみると、図3のようになり、最も細かい文字でも十分判読できる事がわかる。

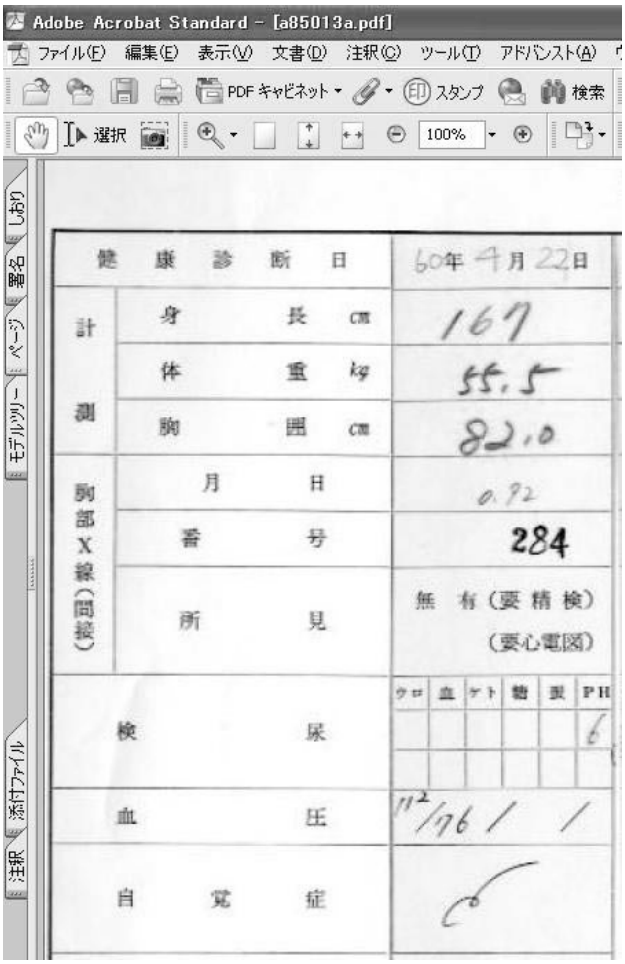

図3 画像ファイルの拡大表示

4. 完全性

e-文書法において完全性は、真実を確保するための要件であり、「電子署名やタイムスタンプなどの偽造不能な署名を電子文書に付すこと」でそれを実現することができる。

タイムスタンプの場合、これを実現するには、第 3 者の認証業者と契約をしなければならない。完全性を保証したいデジタルファイルから、ある方法でハッシュ値（16 バイトまたは 20 バイトの文字列）を計算し、それをタイムスタンプ局に送り、その値を基にして暗号化されたタイムスタンプを受け取ることができる。そのタイムスタンプをデジタルファイルに付加してファイルを保存する。ファイルが

改ざんされた場合、ハッシュ値が異なるので、認証ソフトでファイルの正当性や改ざんの有無を指摘できる。つまり「ある日付以前にそのファイルが存在しており、現在まで改ざんされていない」と言うことを第3者である認証業者に保証してもらうのである。

そのためのコストに関しては、たとえばAdobePDF形式ファイルについて、アドビシステムズ株式会社とアマノ株式会社が、e-文書法対応のタイムスタンプシステム「e-timing EVIDENCE 3161 for Acrobat」を提供している[文献4]。そのサービス料金体系を見ると、従量制メニューとして、1アカウントに関して、初期導入費用6,300円、ランニング費用月額8,925円(1アカウント管理費と1000スタンプ分の利用料を含む)がある。また1000スタンプを超えた場合は、1スタンプあたり8円である。

保存すべき電子ファイルに対して、他社のサービス価格体系を調べても、1 スタンプあたり約 8 円から 10 円程度のコストがかかる計算になるが、国税関係書類の場合、法令解釈通達により、「スキャナー読み取り日が特定できるように電磁記録をまとめてタイムスタンプを付したり、台紙に複数枚の書類を貼り付けて、まとめてタイムスタンプを付したりできる」となっている[文献3]。複数ページのPDF文書に1つのタイムスタンプが許されるということで、タイムスタンプコスト(完全性保証コスト)を大幅に低減する可能性がある。

5. 機密性

機密性は、「文書の盗難、漏洩、盗み見などの防止」[文献3]ということで、デジタル文書やデジタル化文書の保管場所やアクセス場所を管理することである。

これには、まずデジタル文書等の保管場所やアクセス場所への立ち入りを管理しなければならない。ハード的には、鍵の管理やIDカードの管理、閲覧時間の管理などを行わなければならない。また保管や閲覧作業が、インターネットを介して行われる現状では、ソフト的には、アクセス制御や暗号化、アクセス履歴管理、保管の2重化や定期的な照合などが必要である。

また、自社でこのような機密性の高いシステムを構築し管理できないのであれば、信頼の置ける第3者と契約して、管理のすべてを任せるのもひとつの方法である。ここでもコストの問題が出てくるので、保管すべきデジタル文書等の価値とシステムの費用対効果を考えて判断すべきである。

デジタル文書等を責任を持って外部システムに保管してくれるのは、ホスティング業者である。「ホスティングサービス料金」というキーワードで検索すると、いくつかのホスティング業者が出てくる、たとえば120GB(CD約187枚)のスペースを借りると、月額1,500円から4,000円程度の料金、200GB(CD約312枚)のスペースでは月額3,000円から8,000円程度の料金であることがわかる。

また、紙文書で保管する場合、社内の保管場所が満杯の場合は、外に倉庫やトランクルームを借りてそこに搬入し保管することになる。保管コストは地域によって異なるので、「トランクルームサービス料金　岐阜」というキーワードで検索すると岐阜エリアにおけるいくつかのサービス企業が出てくる。たとえばトランクルームの場合、岐阜エリアでは、1.5 帖から 7.4 帖のスペースのトランクルームが月額 7,000 円から 27,000 円程度で借りることができる。月額保管料金以外に搬入の頻度と量を考慮した搬送コストも考えに入れておかなければならない。

6. 検索性

検索性は、「必要に応じて求める文書を探し出すことができること」[文献3]である。文書に対して、キーワードを設定し、それによって文書を検索し、画面に表示し、必要に応じて文書を印刷できなければならない。

学生健康診断票に対しては、「どのようなファイル形式とするか」および「ファイル名をどのように決めるか」、「検索キーを何にするか」を決めなければならない。

文献1では、検索し画面に表示された画像の操作性から、ファイル形式は PDF が有利であるとしている。PDF ファイルは情報交換ファイルとして業界標準となっており、画面表示においては、HTML 文書と同様に、画面クリックによって WEB ブラウザの画面の中に表示できることも有利な点である。また、Acrobat の「ナビゲーションタブ」から各頁にあらかじめ移動先名を付けておけば、PDF 文書の途中頁を HTML 文書からリンクして表示させることもできる[文献5]。

ファイル名については、学籍番号が台紙と対応をとることのできる候補であることから、学籍番号を含むファイル名にすることとした。台紙の表と裏とを区別し、さらに1人の学生が複数の台紙を持つこともあることから、ファイル名は、区分記号（アルファベット1桁）と学籍番号（数字数桁）と付加記号（アルファベット1桁）をつなげたものとした。

区分記号は、学部、大学院、専門学校で、学生健康診断票に含まれる学籍番号のコード管理基準が異なると考えられるので、それらを区別する1桁のアルファベットである。具体的には学部を「a」、大学院を「b」、専門学校を「c」とした。

付加記号は複数枚に対応するもので、台紙の表を「a」、裏を「b」とし、同じ学生の台紙が続く場合は、表裏という順序で「c」、「d」などと続けて付けていくことにした。たとえば図1の台紙のファイル名は、表であるから、「a85107a」となり、図2が同じ学生の裏であれば、そのファイル名は、「a85107b」となる。

検索キーには、まず学籍番号が考えられる。学籍番号はファイル名の一部となっているので、これをキーとした検索は容易である。そのほかにも「姓名」、「生年月日」、「性別」などのキーが考えられるが、台紙を見ながらそれぞれデータを起こさなければならないので、そのための特別の作業が必要である。それに対して学籍番号は、すでにファイル名として設定するよう決めたので、ファイル名を付加するときに同時にデータを起こすことができる。

7. 学生健康診断票検索画面

画像化された学生健康診断票の検索画面は、トップ画面[図4参照]に各冊子の年度と内容情報が表示され、そこからクリックによって各冊子の表示画面に行く事ができる[図5参照]。各冊子の表示は、フレーム表示となっており、左のフレームにある学籍番号をクリックすることにより、右のフレームにその学生の台紙を表示させることができる。いちいち画面を戻さなくても、同じ冊子内にある別の学生であれば、左のフレームをクリックして、次々に右のフレームにその台紙を表示させることができる。

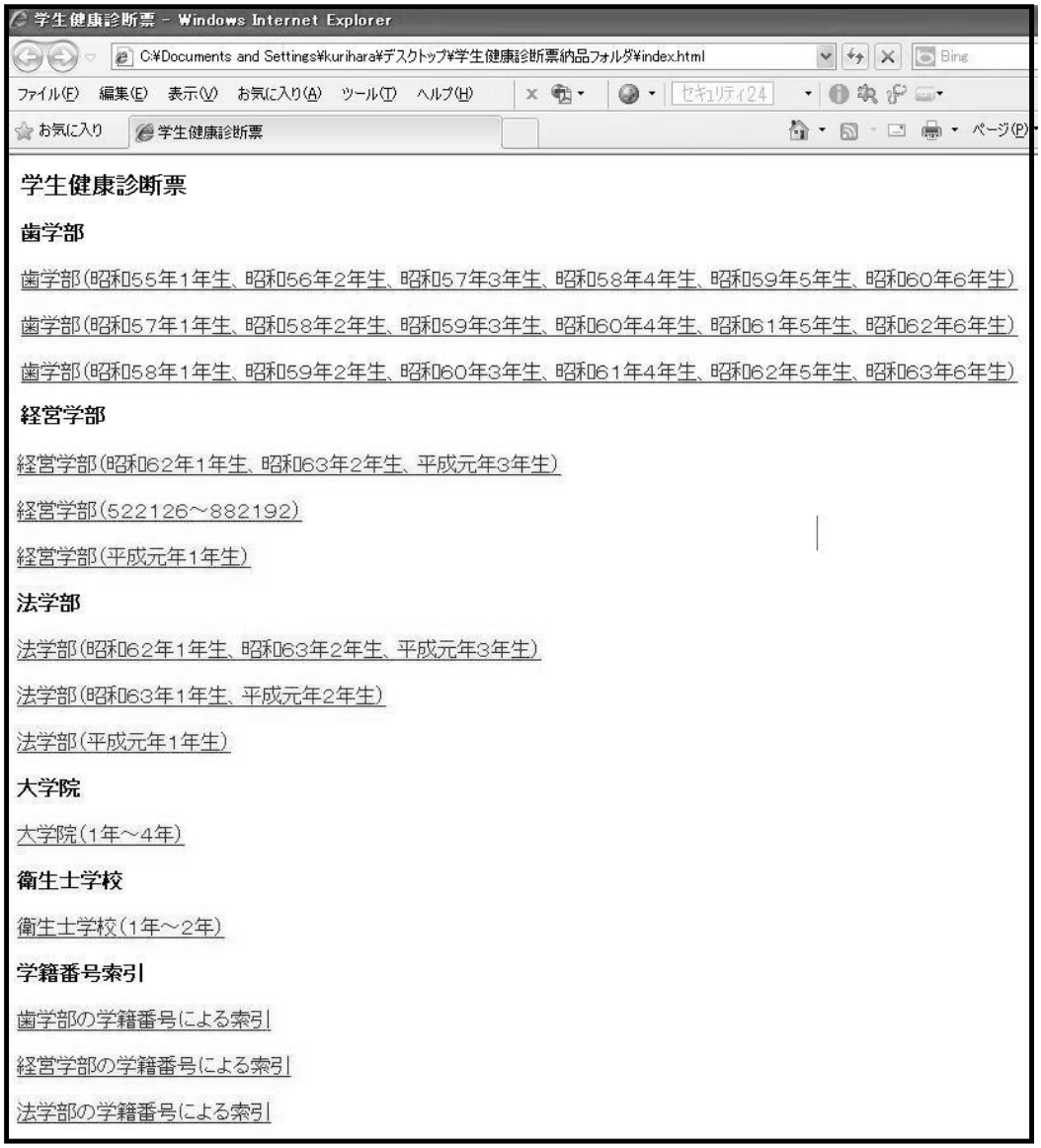

図4　学生健康診断票検索トップ画面

またトップ画面の下部に学籍番号索引があり、クリックによって学部ごとの学籍番号索引[図6参照]に行くことができる。学籍番号索引の学籍番号をクリックすることにより、その学生の台紙を見る事ができる。表示は単純で、図5の冊子表示画面の右のフレームにあるような台紙が全画面に表示されるようになっている。この場合検索は個別検索となるので、別の学籍番号を探す場合は、学籍番号検索画面に戻らなければならない。

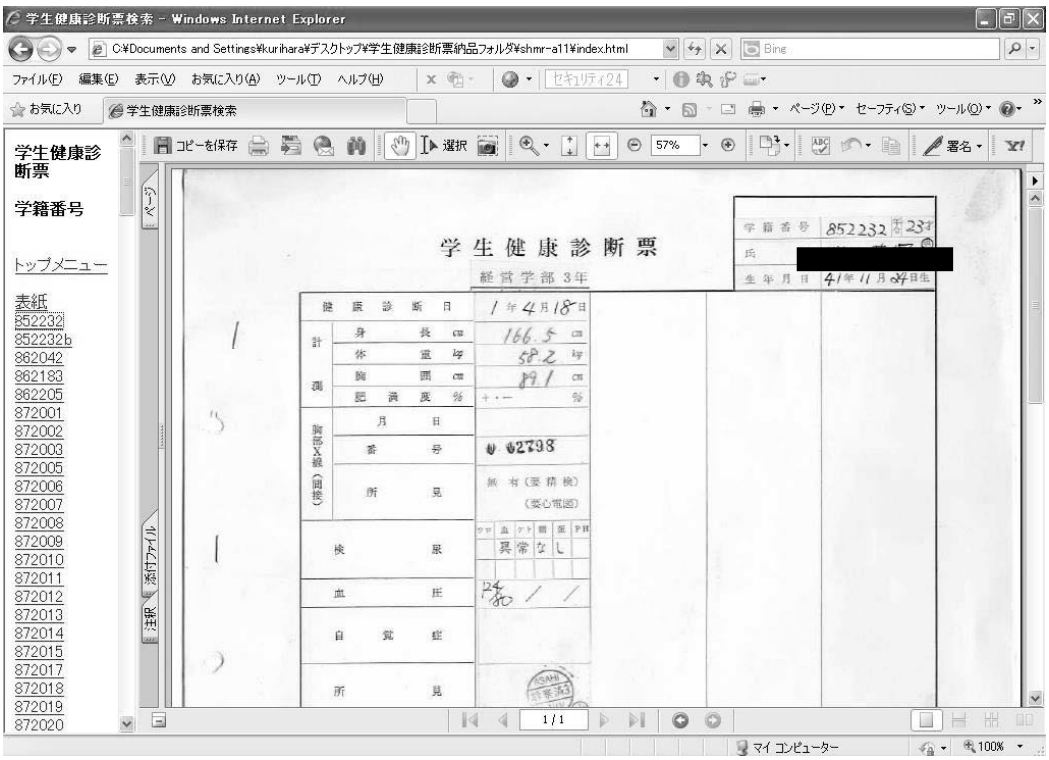

図5　冊子表示画面

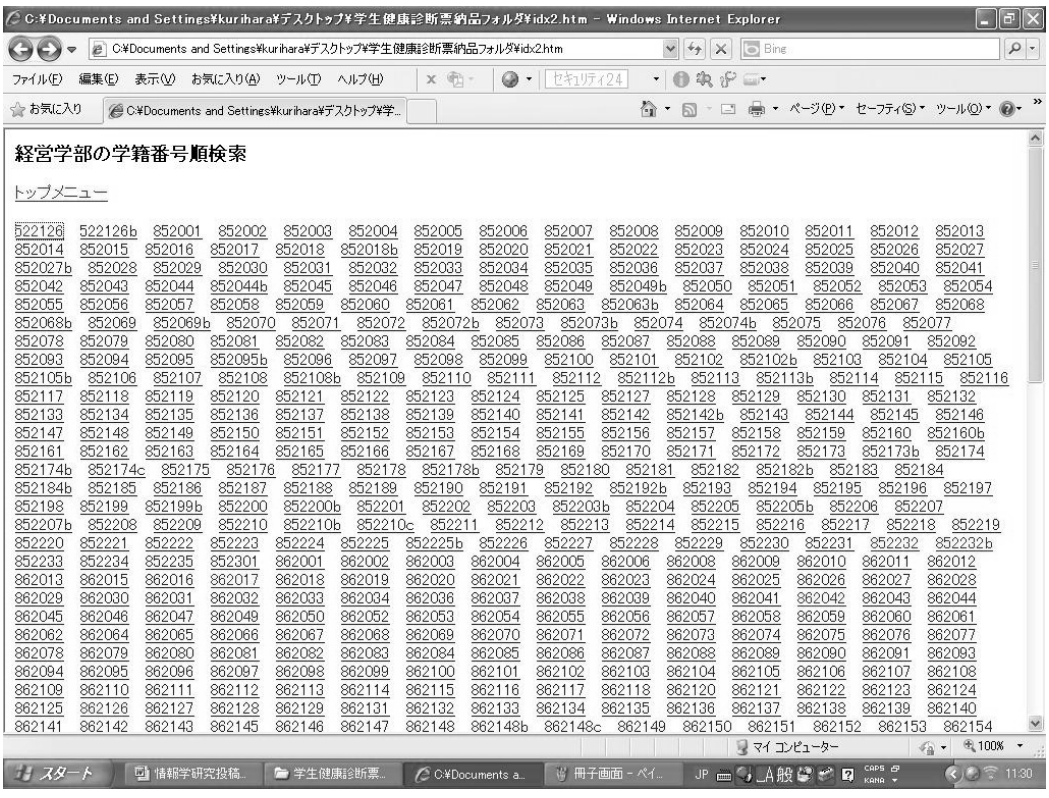

図6　学籍番号索引

7. ファイル名の変更と検索タグの作成

スキャナーにより画像化されたファイルは、記号と連番によるスキャナー独自のファイル名がつくので、「6. 検索性」のところで述べたファイル名設定基準に基づいて、ファイル名の付け替え作業を行わなければならない。

スキャナーによっては直接pdfファイルを出力できず、bmpファイルやjpgファイルを出力する場合があり、これらをpdfファイルに変換する前処理が必要である。そのためのフリーの変換ソフトは、ネットからいくつか探すことができる。

このように多くの画像を検索する HTML 文書は、その画像の数だけ画像にリンクする行を作成しなければならない。今回はファイル名を変更する作業があるので、そのとき同時に検索に利用する左フレームの HTML 文書の行を効率的に作って行くことができた。ファイル名がたとえば「a85107a」で、これにクリッカブル文字列「85107」を付ける場合の HTML 文書の行は、

<hr />85107

である。ここで frame2 は右フレームの名前であり、下線部分はファイル名に応じた可変部分である。ファイル名を修正するときに入力する「a85107a」の一部をとってクリッカブル文字列「85107」を作り出し、その他の固定文字列とこれらを連接することにより、この1行ができる。

連接は、編集ソフトの置換などでもできるが、今回は Excel の連接を使った。ファイル名変更と HTML 文書の行を同時に作成する作業画面を図7に示す。

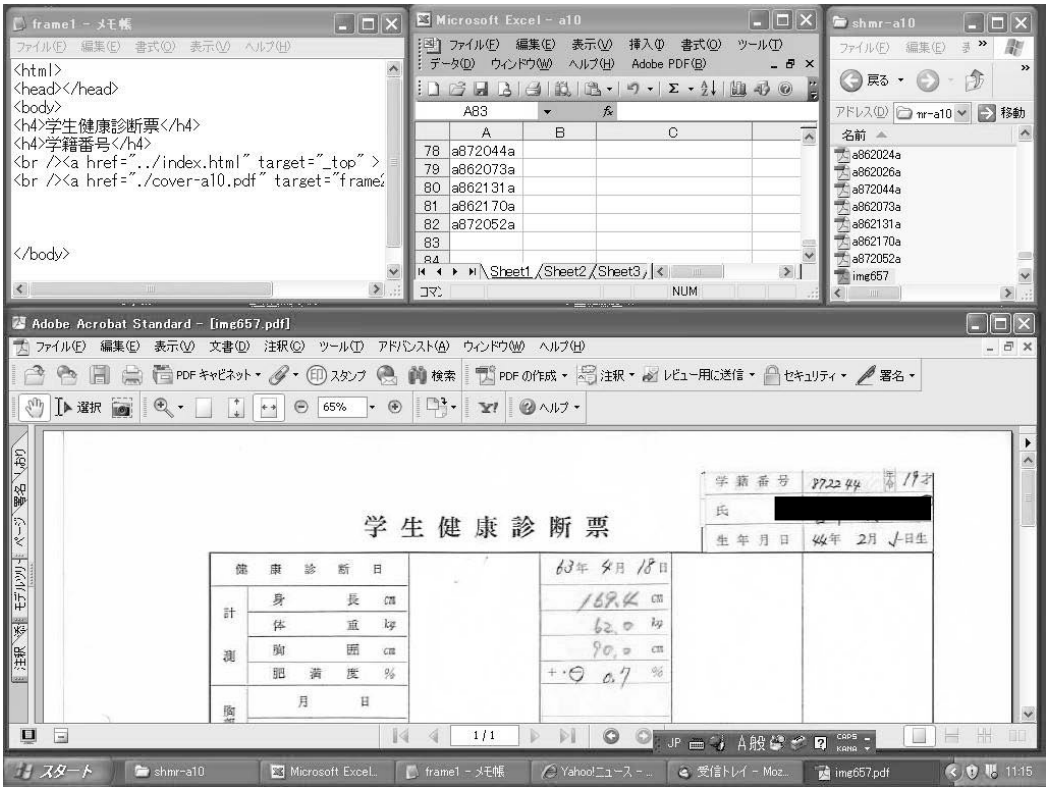

図7 ファイル名変更と HTML 文書の行を作成する作業画面

図7の上の右画面は、Explorer の画面である。下の画面は Acrobat Reader、上の中の画面は Excel、上の左の画面は Notepad（メモ帳）である。

作業は次のような順序となる。まず右上画面の Explorer で、スキャナーで付けられた img657.pdf というファイルをクリックして、下の Acrobat Reader 画面に画像を表示させ学籍番号（たとえば 85107）を確認する。これを Explorer のメニューで、たとえば a85107a.pdf という名前に変更する。このファイル名を Excel のA列に入力する。第1行目を対象行とすれば、Excel のB列には、

=mid(a1,2,5)

という式を入れておけば、ここにクリッカブル文字列 85107 が作成できる。またC列に

<hr /><a href=”./

という固定データ、D列に

.pdf” target=”frame2”>

という固定データ、E列に

という固定データを設定しておき、F列に

=C1&A1&D1&B1&E1

と設定しておけば、F列に

<hr />85107

ができる。1つの冊子すべての画像についてファイル名を変更したあと、F列すべてをコピーして左上のメモ帳の該当箇所に貼り付ければ、目的の HTML 文書が完成する。

姓名、生年月日、性別などさまざまなキーに対して検索を行わせるためには、キー項目のリストから database を作成し、その検索機能を用いたプログラムによってそれを実現しなければならない。今回は HTML 文書を使って、学籍番号から検索する最も単純な方式で検索を実現した。

8. おわりに

この作業の結果、図8に示すように、横幅 90cm3 段のファイルキャビネットいっぱいに入っていた台紙つづり19分冊がCD1枚（容量640MB）に十分格納できた。またその容量は図9に示すよう、19 分冊で約 260MBであり、そのCDも4割強しか使われていない。

この激しい減量と減容の経験が、本作業から得られた結果である。すでに存在している電子書類の原本（ワード文書やエクセル文書などのオフィス文書）を PDF 化してキーを付加し検索できるようにすることは、情報共有化の点で重要なことである。また今回行ったように、原本が紙文書のものでも、スキャナーで画像化してデジタル化文書にする利点は明らかである。

組織の中に保管されている紙文書の量は膨大であるが、デジタル化に対して費用対効果を考える上で本報告がひとつの参考になることを希望している。

図8　学生健康診断票のキャビネット（文献1より転載）

図9　学生健康診断票フォルダの容量

参考文献

[1] 吉田晴人、森嶋孝太、松森絵美、野口良祐、吉田祐一「紙文書の画像化と検索に関する一考察」、2010 年度経営学部情報管理学科卒業論文

[2] ＠IT情報用語辞典「e-文書法」、http://www.atmarkit.co.jp/aig/04biz/ebunshoho.html、アイティメディア株式会社

[3] 横山三郎（税理士）、http://www.atmarkit.co.jp/aig/04biz/ebunshoho.html

[4] タイムスタンプの概要／アマノビジネスソリューションズ株式会社、http://www.e-timing.ne.jp/tsa/service/summary.html

[5] HTML ページから PDF ファイルの特定の場所へのリンクを作成する方法（Acrobat 7.0）、http://kb2.adobe.com/jp/cps/226/226119.html

栗原　和夫（経営学部情報管理学科教授）