鹿児島純心女子短期大学研究紀要　第31号，245－257　2001

学内出版物の電子化保存形式に関する一考察

吉田　尚史

A Consideration on Digital File Formats Converted from In-School Publications.

Naofumi Yoshida

現在，「電子図書館」に関して様々な取り組みがなされているが，実際に稼働しているシステムとしては書誌情報，所蔵情報の提示がほとんどであり，本文を含めた自由な参照はいまだ一般的に可能ではない。これは著作権など内容に関する諸権利の問題および著作物を利用する場合の利用料の問題，公開する場合の技術的な問題，実現するための経費の問題など，様々な問題が存在するためである。

本稿では電子文書の公開にあたり，学内・図書館で出版される紀要・館報を電子化する場合に考慮されるべきファイル形式の選択を問題とし，その作成に当たって考慮すべき点，公開された文書のセキュリティの問題，検索に対する特性などを比較し考察を行う。

Key words：［電子図書館］［学内出版物］［文書の電子化］［マークアップ］［pdf］

(Received November 6 , 2000)

１．はじめに

現在，日本国内の４年制大学･短期大学等高等教育機関の付属図書館において，インターネットへ接続し所蔵資料の書誌情報，所蔵情報等の２次資料を公開しているものは多数にのぼる[1]。しかしながら所蔵資料の内容までネットワーク経由で参照できるシステムはごくわずかである。

電子図書館を定義する視点としては次のような点が指摘される[2]。

１）ライブラリ・オートメーション

図書館業務の機械化，情報化の推進。電子目録作りや購入貸出し業務の自動化。図書館相互のオンライン情報交換など。

２）コンテンツのデジタル化

書籍の内容のデジタル・データへの置き換え。

３）インターネット・ライブラリ

インターネット上の情報資産を体系的にリンクさせ，サイバースペースの中に仮想的な図書館・博物館を構築。

鹿児島純心女子短期大学生活学科生活学専攻生活経営コース（〒890-8525 鹿児島市唐湊４丁目22番１号）

また別の観点として「電子図書館とは，著者名，書名，主題等による必要情報の検索，資料の取り出し，必要な部分の入手までのすべてが電子化されている図書館と考えることができる」[3]という指摘もある。

これらの観点にたってみれば現在のシステムは上記のような「電子図書館」という定義とはほど遠い状況である。この背景には，保存・公開する際の著作権等諸権利の問題，あるいは参照される際に発生する内容の利用料に関する問題，保存・公開する際の技術，あるいはそのような技術を実現するための経費というような問題が存在する。

収蔵図書の著作権などに関連する問題への解の一つはインターネット上に構築された図書館である『青空文庫』[4]が示している。これは死後50年経過し作品の著作権が消滅した作品を電子テキスト化し保存・公開している。また著作権が有効な作品については，青空文庫自身が保存公開するのではなく，著作権者自身が電子化されたテキストを自らのサーバースペース内に置き，それに対して青空文庫の図書カードからリンクする形を取っている。

以上のような状況に鑑みて，本稿においては学内で出版される紀要，図書館報などのテキスト類を電子化し，公開するという状況を前提とする。これはそのようなテキストについて著作権を保有するものが各図書館にとって限定されるため，公開の了承を得る様な場合に著作権の問題をクリアすることが容易であるということが理由の一つである。

さらに重要な理由として，このようなテキスト類が電子図書館の特性そのものに大きな意味を持つことが上げられる。電子図書館の特性として，インターネット上に非限定的に公開可能なデータであれば，それを各館が重複して保存しておく必要がないことが指摘できる。現実社会における図書館は，各館がそれぞれ物理的な書籍を所蔵する必要がある。もちろん所蔵していない資料についても各館の利用協力により他館資料を参照することは可能であるが，他の図書館を利用者が直接訪問する，あるいは他館からその資料や資料の複写を郵送する，などの労力と時間・経費が必要となる。一方いわゆる電子図書館であれば，他の館に保存されている資料で，非限定的に参照が許されるものであれば，いながらにしてその資料を参照することが可能である。他館が所蔵し公開的な参照が許されている資料との資料を除けば，自館で電子化し所蔵すべき資料は限定的なものとなる。すなわち⑴全ての電子図書館が全て横断的に検索可能であり，⑵各館の資料がどの館からも公開的・非限定的に参照可能であれば，１個の巨大な図書館として機能することができる[5]。これが学内テキスト類の電子化を考察することの重要な意味である。

もう一つの大きな問題として，テキストの電子化，その保存および検索にかかわる技術的な問題がある。電子図書館という語を，インターネットを利用する，あるいはローカルなネットワーク内のみで利用する，もしくは単独のコンピュータ内であれ，テキストを電子化し，保存および公開し，検索，参照するための技術とその技術によって提供されるサービスと定義すれば，そこには⑴テキストの電子化，⑵電子化したデータの保存方法，⑶データの公開方法－検索と提示・参照の技術，⑷著作権保護のための技術，⑸データの授受に利用される技術等が，検討の対象としてあげられる。

本稿では特に⑵の電子化したテキストを保存する場合どのような形式が利用可能かということを考察のテーマとするが，これは⑶のどのようにそのデータを提供するかということ，およ

び(5)のデータの授受をどのように実現するかということと密接に関連している。従ってこれらの点とともに考察される必要があるが，本稿では，データの授受については TCP/IP をプロトコルとするネットワーク上の WWW サーバーとクライアントの組み合わせたシステムを，またデータの公開・提供の方法としてはサーバー側でのデータの蓄積・検索・提示を行うシステムを前提とする。これらは実際に電子図書館的なシステムを構築する場合に広く利用されている方法であり，現実として一般的であるためである。また文中の「ローカル (側)」という語は利用者が操作しているコンピュータを指す。

2．電子テキストの保存形式の概要

2－1．画像ファイル形式

これは一度印刷されたテキストをスキャナやデジタルカメラを利用して書面を画像の形式でファイルに保存する方法である。ファイルの保存形式としては WWW を利用してデータの参照が行われる場合，ブラウザが一般的にサポートする GIF (Graphics Interchange Format), JPEG (Joint Photographic Coding Experts Group) 形式が利用される。前者はアメリカのパソコン通信ネットワークである CompuServe で用いられていた画像データの方式で，256色までの画像を効率よく圧縮し，すなわちファイルサイズを少なくし保存することができる。後者はフルカラーの画像を圧縮保存するための方式であり，写真などのデータを効率よく保存することが可能である。

この形式は WWW で一般的に用いられている画像フォーマットであるため，HTML 内でファイル名を指定することにより，作業を中断することなくブラウザ内で参照することがかのうである。またローカル側でファイルを保存することで，一般的なグラフィックソフトウェアを用いてファイルを開くこともできる。

この形式を用いて電子化された文書を公開している例として，図書館情報大学付属図書館がそのホームページで図書館報を公開している。図書館情報大学付属図書館では，1985年のVol. 1以来全ての図書館報の目次を HTML 形式で公開しているが，Vol.10（1994）以降は HTML ファイルによる全文テキストと GIF 形式による画像データの両方を公開している。さらに Vol.20（2000）以降は PDF 形式のファイルも同時に公開されている。

この方式では，印刷されたものをそのまま保存するため，印刷時のレイアウトや書体などをそのままの形で参照することが可能である。元のレイアウトや装丁が重要な意味を持つ場合など，画像ファイルとしての保存が意味を持つ場合も多い。慶応大学で進められいている「HUMI プロジェクト」[6]では，収蔵しているグーテンベルク42行聖書などの稀覯本をデジタル化しているが，この公開に当たって，JPEG 形式のフォーマットが用いられている。

電子化された原稿がない場合，あるいは印刷されたものを OCR や人間の手作業によって電子化することが困難な場合などもこの方式を用いる理由となる。印刷物の全てのページをスキャナなどを用いて画像ファイルとして保存する作業は容易なものではないが，物理的な印刷物から文字データをデジタル化する作業に比べれば時間や労力の点で有利な方式である。

また先述のように「コピー」と「張り付け」による引用が不可能ということは，それを再利用することが容易ではないという意味で,当該文書の諸権利の保護にとって利点となるだろう。

内容の検索については，文書内の文字データが文字として処理されるのではなく，画像の一部として処理されるため，文書内のどのページにどの様な文字データが含まれるかを検索するのは困難である。従って当該文書のファイルに対する検索および複数ファイルに対する検索も閲覧用のソフトウェアを用いるだけでは不可能である。

また画像として保存する際の解像度などの指定によりファイルのサイズが膨大になる可能性がある。また参照する側がこのテキストに含まれる文字データの一部を引用するためには，他のファイル形式のように画面上でコピーした文字データをそのまま同じ画面上で自らのテキストに張り付けるという操作は不可能である。

2－2．マークアップランゲージテキストファイル (SGML/HTML/XML)

一般に文字データだけから構成される電子ファイルをテキストファイルという。これはアルファベットや記号，ひらがな，カタカナ，漢字などの情報を表記できる一方，ページや文字などの書式情報などを保持することはできない。また図表などのデータを保存することも不可能である。

WWW上で公開する文書を記述するための HTML (Hyper Text Markup Language) 形式の文書はこのテキストファイルである。これは通常の電子化された文章にタグを挿入することにより，文章の体裁，構造を記述し，画像や他のテキストへのリンクなどをその文章に付加することができる。HTML を用いて記述された文章はタグを含めて文字データのみで構成されたテキストファイルであるが，その中に画像や音など文字以外のデータが保存されたファイルを参照するタグを含めることで，HTML を閲覧するためのソフトウェア（ブラウザ）により，そのページの中に文字データ以外の情報を含めて提示することが可能である。

HTML は SGML をベースとしており,インターネット上の WWW サービスを利用して文書を公開･参照する目的で開発されたものである。HTML は文書の構造をマークアップするのではなく文書のレイアウトや他文書へのリンクを記述するなどの目的のタグセットとなっている。

SGML (Standard Generalized Markup Language) は当初，電子組版のために考案されたものであった。これは文章の構造を記述するための言語仕様であり，特定の目的に添ってそれぞれタグを定義し，そのタグを文書内の要素に付加していくことで文書内の構造を表現するためのものである。SGML 形式の文書がタグによって文書構造を表現し，テキストファイルで保存されているということは，タグセットがどのような内容を示しているかさえ明らかであれば，その文書の内容を特定のソフトウェアによらずに容易に解析することができる。すなわちコンピュータ間でのデータの授受における互換性が非常に高く,このような意味で SGML は電子文書の管理において重要な意味を持つものとして各所で利用されることとなった。

しかしながら SGML は HTML のようにページの体裁あるいは図表などを表示するための記述などを目的とするものではなく,また HTML や XML のようにインターネットを利用した文

書の公開･参照を前提としたものではないため，サーバーに保存してある SGML 文書を一般的な HTML ブラウザを通して要求し，参照するという利用の仕方は困難である。何らかの形でHTMLに変換し利用者へ提示する必要がある。

タグをマークアップすることでデータの構造を表現するという思想は XML (eXtensible Markup Language)[7] も同様である。これは SGML と同様に HTML のように文章を直接マークアップするためのタグセットではなく，それらのマークアップ言語を規定するためのメタ言語である。すなわち XML の仕様に基づき様々な内容を表現するための言語仕様が作られ(eg. MathML, XHTML etc..)，それに基づき実際のマークアップがなされる。

SGML・XML を利用したマークアップでは，文書の内容に即して文書構造定義 (DTD:Document Type Definition) を作成し，それによって文書の構造を明示する。実際に表示する際のレイアウトは，CSS (Cascading Style Sheet)･XSL といったスタイルシートを作成し，この中に細かな表示上の書式を指定することで実現される。

SGML 形式で記述された電子文書を閲覧するためのソフトウェアは現在のところ広く普及してはおらず，一般的に普及している HTML ブラウザ，例えば Microsoft 社の Internet Explorer や Netscape 社の Navigator/Communicator といった HTML ブラウザでは書式情報などを含めて表示することができない。これらの形式の文書を見るためには，参照要求があった時点でサーバー側で HTML 形式に変換して提示する必要がある。XML の場合には「Internet Explorer」の Ver. 5 以降で XML の表示に対応しており，当該 XML 文書内でスタイルシートを指定する記述があればこれをレイアウトまで表示することができる。

これらの形式の文書はテキストファイルであるため，ローカル側での検索，サーバー側での検索のどちらにも対応できる。サーバー側で複数の文書内容を特定の文字列で検索し，それが含まれている文書を提示することが可能である。また利用者が現在ブラウザで表示されている一つの文書の中で特定の文字を検索することも可能である。

2－3．PDF 形式

PDF （Portable Document Format) 形式は Adobe社 によって開発された電子文書用のファイルフォーマットである。DTP ソフトウェアやワードプロセッサソフトなどアプリケーションでレイアウトした文字や画像を含んだ文書を，紙に印刷したものと同じ体裁でディスプレイ上に表示可能な形式に変換することができる。さらに PDF 形式保存した場合，ネットワークで配布することを前提としているため作成したアプリケーションで保存した場合よりファイルサイズを小さくすることが可能である。また文書内で利用しているフォントを文書内に埋め込むことができるため，どのような環境でその文書が参照されていたとしても作成した際のレイアウトを忠実に再現することができる。

PDF 形式のファイルを参照するためのソフトウェアである「Acrobat Reader」は無償で配布され，各種の PC ハードウェアおよびオペレーションシステム (OS) に対応しており，また HTML ブラウザのプラグインとしても機能するため，HTML 文書からリンクを張ることで容易に参照できる。また PDF 形式を作成するためには，専用のソフトウェアである「Acrobat」

を用いるが，これは原稿が電子文書となっていることが前提となっており，これを作成するための DTP ソフトウェアやワードプロセッサなどが必要である。

さらに PDF 形式は電子文書を作成する側が，その文書を参照するものにどのような参照の仕方を許可するか，あるいは許可しないかを作成の際に細かく指定することができる。これにはファイルを開くためとセキュリティオプションを変更するためのそれぞれのパスワードを指定することや，画面に表示されている文書の印刷の可否，文書変更の可否，ファイル内に含まれる文字データの選択の可否，これはすなわち文字データを他のソフトウェアにコピー可能かどうかと同じ意味であるが，これらの点について電子文書を作成する側が細かく指定することができる。

これらのセキュリティオプションは，参照する側が画面でその文書を読む以外に全くないようにふれることを不可能にし，かつその文書の内容を他の電子文書に引用しようとする場合には，入力しなおす以外の方法を不可能とするようにも指定できる。またファイル自体にパスワードを指定することで，特定の条件を満たしたもののみに文書を参照するようにすることも可能である。

内容の全文検索については，上述の「Acrobat Reader」を用いれば，一つの文書内に含まれる特定の単語あるいは字句で検索を行うことができる。保存された複数の PDF 文書に対して検索を行うためには，PDF 形式の文書を作成するためのアプリケーションである「Adobe Acrobat」が検索を実行するコンピュータにインストールされている必要がある。PDF 形式を用いた電子文書の公開は非常に多くの事例が見られる。特に白書や報告書など政府系文書や，本稿で取り上げている学内出版物など，一般書店での入手が困難であり，かつ収蔵する図書館が限られるような文書も多くこの形式で公開されている[8]。

2－4．MS-Word 文書形式

Microsoft 社の製品である，「Word」(以下 MS-Word) は現在利用されているワードプロセッサソフトとしては非常に大きなシェアを持つものである。WWW を利用して公開する文書の形式としては一般的ではないが，政府による報告書などがこの文書形式で公開されている場合がある。

Word 形式で文書を保存・公開することの利点としては，そのために特別な操作・処理を必要としないことである。元の原稿がこの形式で保存されていれば，そのままの形で提供することができる。他の形式のように，元々の電子化された保存形式からの変換やタグをつけるという作業が必要でない。また DTP 専用のソフトウェア程ではないにしろ，ある程度，作成者の意図するレイアウトなどの書式情報を保ったまま，利用者に提示することができる。

HTML 文書内からリンクされた Word 形式の文書を参照する場合，ローカルのコンピュータ上に「MS-Word」か，Microsoft 社が無償で公開している参照専用のソフト「Word Viewer」が必要である。

ワードプロセッサのデータファイルであるということは，利用者による内容の参照や印刷だけでなく，ローカル側で編集ができる可能性があるが，「MS-Word」の「文書保護」の機能を

利用することで，文字データの選択・コピーを不可にするなどある程度のセキュリティをかけることができる。

また「MS-Word」の持つ検索機能を利用すれば，ローカルに保存したファイル内で全文検索が可能であるし，OS や他のソフトウェアの機能を利用して複数のファイルに対して特定の文字列を含むものを検索することも可能である。

3．各形式の比較

上記の各保存形式について，1）サーバーにおける検索性，2）ローカル側でのファイルの操作に対する制限，3）電子化の変換作業工程数，の 3 つの点から比較してみる。

3－1．サーバーに保存されている場合の各ファイル形式の検索特性

次に各形式で保存されたファイルに対するサーバーに上での検索についての比較である。

ここではインターネット上のホームページを検索するための Yahoo や goo などの検索サイトを用いて公開した電子文書を検索する場合と，文書を公開している図書館のサイト内で，独自の検索用ページなどを設定し，そのサーバー内で検索を行い結果を表示する場合についてみてみる（表 1）。

表 1

ファイル形式	Yahoo, goo など当該サーバー外の検索エンジンによる検索	Namazu 等検索エンジンを利用した当該サーバー上での検索
画像ファイル	▲	△
テキスト (HTML 等含む)	○	○
PDF	▲	○
ワード	▲	○

△－データベースなどと組み合わせることにより可能
▲－<Keywords>タグの適切な使用により可能

検索サイトが特定の文書に含まれる内容を検索する場合には，通常その文書に含まれる特定の文字列を索引化し，それによって検索語に該当する文書を検索する。この索引化はその文書内の文字データを用いて行われるため，文書の保存形式が画像ファイルである場合には当然そこから索引を作成することはできない。また PDF 形式，ワード形式の場合も，保存形式がプレーンテキスト[9]ではないため同様である。従ってこれらの形式の文書を索引化し検索に該当させるためには，その文書へのインデックスとなる HTML 文書の中で <KEYWORD> タグを用いることで，検索エンジンに内容を参照させることができる。

文書が保存されているサーバー上での検索については，そのサーバーの OS にもよるが通常「Namazu」[10]や「SSE」といった検索エンジンが用いられる。これらの検索エンジンも対象と

なる文書について索引化を行い，ウェッブサーバーを通して検索要求が送信されてきた場合，この索引をもとに検索語に該当した文書の保存場所などの情報を返す。

検索サイトを用いた場合との違いは「Namazu」を利用した場合，HTML 形式以外でも PDF 形式や Word 形式の文書に対しても索引化が可能であるという点である。検索用のフォームを用意し，そこに入力された検索語などの内容を検索用のスクリプトに渡すことで，HTML 形式の文書だけでなく PDF 形式のものや Word 形式のものでも文書内容によって検索を行うことができる。

以上の検索方式においては，いずれも索引化を行い，それを指定された検索語に照会することで，その語が含まれる文書を検索する。従って索引化の際に含まれなかった単語が検索語に指定された場合，また検索語が文章であった場合などは，それらが含まれた文書が保存されていたとしてもそれを見つけだすことができない。このような場合には保存されている文書に対して全文検索を行うことで，目的の文書を検索することが可能である。

画像ファイルのように全く文字データが含まれていない場合であっても，その全内容を OCR などを用いて文字データ化したものをサーバー上のデータベースに保存をしておくことで全文検索が可能となる。鹿児島大学学内研究成果電子化実験[11]では，PDF 形式とともに画像形式でも文書を公開しているが，これに対する検索については上述の方法を用いている。

3－2．ファイル形式による利用者側の制限

サーバーによって公開されている電子文書をローカル側で閲覧あるいは保存した場合，当該ファイルを閲覧あるいは編集可能なアプリケーションを用いた場合の利用者側の制限事項についてついて比較すると下の表 2 のようになる。

表 2

ファイル形式	HTMLブラウザを除いた各ファイル形式の閲覧に必要なアプリケーション	閲覧	文書内容の編集	印刷
画像ファイル	−	○	×	○
テキスト (HTML 等含む)	−	○	○	○
PDF	Acrobat / Acrobat Reader	○	●	●
ワード	MS-Word / WordViewer	○	●	●

●－作成時の指定による

画像ファイルの場合，文書の内容は文字データとして保存されるのではなく，色情報をもった点の集合として保存されており，それを表示する場合には文字として人間が読むことが可能であるが，コンピュータには文字データとして認識されていない。従ってその内容の特定の部分を選択し文字としてコピーする，あるいは内容を書き換えるなどの編集操作を行うことはできない。

PDF 形式，Word 文書形式では，内容の編集および印刷について，文書の作成時に制限を加

えることができる（図１，２）。

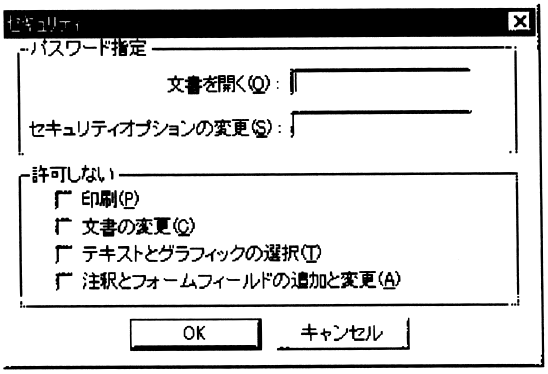

図１　PDF 文書のセキュリティ

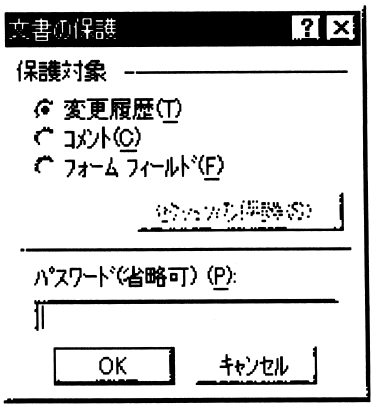

図２　Word 文書のセキュリティ

PDF 形式の方はさらに文書を開く際のパスワードを指定することで文書の参照にも制限をかけることができる。セキュリティオプションの変更については「Acrobat」において「許可しない」項目が指定された場合，利用者が「Acrobat」もしくは「Acrobat Reader」で文書を開いたとき，作成者が指定したパスワードを知りうる利用者のみそれらの項目の制限をはずして編集操作を行うことができる。

Word 形式文書の保護機能は PDF 形式ほど細かな指定項目がないが，保護対象にフォームフィールドを指定することで，「MS-Word」(Internet Explorer内で開いた場合を含む)「Word Viewer」どちらで開いた場合もほぼ全てのメニューが操作できなくなる。そのような場合でも印刷に関するメニュー項目は操作可能であるため，この機能を用いて印刷に関する制限をかけることはできない。

これらの機能を利用することで公開された文書に対する編集や印刷を制限できるが，これは利用者に対して次のような選択肢を提示できることを意味する。すなわち，画面上でその文書を閲覧することのみ許可するか，紙への印刷を許可するか，あるいは内容をコピーし利用者が編集中の他の文書への引用を許可するか，PDF 形式，ワード形式の場合には編集用のアプリケーションが利用できる場合は内容の改変までを認めるか，といった点である。

また PDF 形式では文書の参照自体，パスワードを付加することで制限できるが，これは特定の利用者のみに参照を許す場合に有効である。他の形式で保存する場合でも，サーバー上の WWW サーバーを適切に設定することで，特定のページへの参照を制限することが可能である。

３－３．電子化の変換作業工程数

最後に文書を電子化し各ファイル形式に変換する際に必要となる作業工程数について比較してみる。表３ 「画像ファイル形式」～表６ 「Word 形式」の各行の項目は，それぞれ電子化

の対象となる原稿や文書が作業を始める段階でどのような形態をとっているかを示している。項目中に「(ファイル無し)」とあるのは，紙に印刷された文書を電子化する際に，元となる電子化ファイルが存在しないことを意味している。また「出版済み原稿」とは最終的な形態で印刷所などを利用し，製本まで済んだ形態の原稿の場合を，「印刷済み原稿」とはそれ以前のプリンタで印刷したものを指す。

各形態から特定の保存形式への作業工程としては各列の項目の段階を経ることを仮定する。表7は各形式の作業工程数をまとめたものである。

表3　画像ファイル形式

	OCR/原稿入力	書式情報の付与	印　　刷	スキャナによる読みとりと保存
手書き原稿	○	○	○	○
印刷済み原稿(ファイル無し)				○
書式なしで保存	−	○	○	○
書式付きで保存	−	−	○	○
出版済み原稿(ファイル無し)	−			○

表4　マークアップランゲージ

	OCR/入力による電子化	タグの付与
手書き原稿	○	○
印刷済み原稿(ファイル無し)	○	○
書式なしで保存	−	○
書式付きで保存	−	○
出版済み原稿(ファイル無し)	○	○

表5　PDF形式

	OCR/入力による電子化	書式情報の付与	Acrobatによる変換
手書き原稿	○	○	○
印刷済み原稿(ファイル無し)	○/−	○/−	○
書式なしで保存	−	○	○
書式付きで保存	−	−	○
出版済み原稿(ファイル無し)	○/−	○/−	○

(記号が2つある場合、前者が文字データでPDF化、後者が画像としてPDF化)

表6　Word形式

	OCR/入力による電子化	書式情報の付与
手書き原稿	○	○
印刷済み原稿(ファイル無し)	○	○
書式なしで保存	−	○
書式付きで保存	−	−
出版済み原稿(ファイル無し)	○	○

(「書式付きで保存」はWord 形式か、MS-Word で開ける形式で保存されていると仮定する)

表7

ファイル形式	手書き原稿	印刷済み原稿(ファイル無し)	書式なしで保存	書式付きで保存	出版済み原稿(ファイル無し)
画像ファイル	4	1	3	2	1
テキスト(HTML 等)	2	2	1	1	2
PDF 形式	3	3 / 1	2	1	3 / 1
Word 形式	2	2	1	0	2

(記号が2つある場合、前者が文字データで PDF 化、後者が画像として PDF 化)

PDF 形式の場合にはスキャナなどを用いて印刷物の各ページの画像を PDF ファイルとして保存することができる。この場合には画像ファイルと同様に文書の内容を文字データとして参照・利用することはできない。

単純に工程数からいえば，原稿の電子化する際の編集に「MS-Word」を用い，Word 形式で保存，公開した場合がどのような元原稿の形態であっても工程数を少なくすることが可能である。なお，他の形式を考察する場合との整合性のため Word 形式の場合にも「印刷済み原稿」「出版済み原稿」の項目を加えてあるが，一度印刷されたものを Word 形式に入力，レイアウトし直すことは現実的に考えづらい。

次に工程数が少ないのはマークアップしたテキスト形式である。これは元原稿が電子化されていれば，タグを付加していく作業のみで公開できるためである。しかしながら文書の内容に即してタグを付加する作業は機械的に自動化することが困難であるため，工程数としては少ないが作業量・作業時間が少ないという意味ではない。

4．まとめ

本稿では，大学・短大内で出版されている紀要や図書館報などの学内出版物を電子化しそれを公開する場合，どのようなファイル形式を用いるかということを問題の中心とした。

現実的な作業を行う場合には，電子化する元文書の形態や，サーバーにおける検索方法の構築など，上記のように単純に考えることができない。しかしながら，より一般的，より現実的な方法を模索する際の手がかりとして，次のようなこと指摘できるだろう。

実際の作業に当たっては上記のような工程数の問題とともに作業量・作業時間が大きな問題となる。従って電子化された文書をマークアップする場合には作業の工程としては少ないが実際には膨大な作業を行う必要がある。このようなことを考慮すれば，出版済みの文書から電子化を行うのであれば画像ファイルを，何らかのアプリケーションを用いてレイアウトが済んでいるファイルが存在するのであれば PDF 形式を用いるのが適しているだろう。これらの形式は先述の通り文書内容の保護の点からも有利である。またこれらの形式を用いた場合のサーバー上での検索では PDF の場合には検索エンジンによって対応が可能であるし，画像形式の場合には，文書内容全文を含んだテキストファイルを用いてデータベースを構築することで対応することができる。

工程数を考えれば Word 形式の文書も選択肢として考慮できるが，紙の形態で出版され電子化されたファイルが存在しない場合には，この形式を選択する利点がない。また利用者の側では Word 形式の文書を参照するために専用のソフトウェアを必要とすること，またそのソフトウェアが対応するコンピュータの環境，ネットワークを用いて文書を公開する場合の最終的なファイルサイズの問題，レイアウトの一貫性の保証などの観点からいえば PDF 形式の方がより有利である。

これらのことを考慮した上で，利用者がより利用しやすい環境を考えれば，電子化された文書を自由に検索することが重要になってくるが，これにはサーバー内で検索を行うための様々な「仕組み」が必要となる。この点について本稿でふれることができなかった。これについては機会をあらためてふれる。

註

1：農林水産研究情報センターのホームページにある「日本国内図書館 OPAC リスト (2000年10月23日版)」によれば現在215の組織が OPAC を公開している。http://ss.cc.affrc.go.jp/ric/opac/opaclist.html

2：『電子図書館が見えてきた』，宮井均・市山俊治, 1999，NEC クリエイティブ，pp.38-39.

3：『未来の図書館』，原田勝, 1987，松籟社，p.195

4：http://www.aozora.gr.jp/
このほかに同様のサイトとして,「日本文学関係テキストファイル等（作品別･五十音順）」(http://www.konan-wu.ac.jp/~Ekikuchi/linkd.html),「電子図書館」(http://www.wao.or.jp/naniuji/04bekkan.htm) などがある。また世界規模で見れば最も有名なものとして「プロジェクトグーテンベルク」(http://promo.net/pg/) がある。

5：九州大学の OPAC 横断検索がこの道筋を付けるものとの指摘もある。『電子図書館が見えてきた』，宮井均・市山俊治，前掲，p35

6：http://www.humi.keio.ac.jp/

7：http://www.w３.org/XML/

8：図書館情報大学ホームページ内の「全国の図書館報のページ」(http://www.ulis.ac.jp/library/Jpnpg/gate-kanpo.html)からリンクされている九州内の大学７校の内，館報の公開が確認できたものが５校，その中で PDF 形式を用いて図書館報を公開しているものは４校であった。

9：本稿では，可読文字のみによって構成されているファイルをテキストファイルと呼び，特に自然言語を用いて内容を記述されたものをプレーンテキストと呼ぶ。

10：http://www.namazu.org/

11：http://websv.lib.kagoshima-u.ac.jp/pdf/kiyotop.html